Interpretation of images from intensity, texture and geometry

Vestergaard, Jacob Schack

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Interpretation of images from intensity, texture and geometry

Jacob Schack Vestergaard

Kongens Lyngby 2014
PhD-2014-346

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Building 324, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031
compute@compute.dtu.dk
www.compute.dtu.dk PhD-2014-346

Summary (English)

The goal of the thesis is to develop flexible mathematical methods for quantitative interpretation of image content. Problems from research areas as diverse as evolutionary biology, remote sensing and materials science have motivated the methodological development. The solutions are inspired by classical mathematical image analysis techniques, information theory, probabilistic graphical models and manifold learning.

Specifically, the thesis revolves around describing three major components of images, namely intensity, texture and geometry. Intensity distribution modelling is important for obtaining useful global representations of the raw image data. Texture description provides a local representation of the image content, useful for descriptive and discriminative scenarios. Geometrical knowledge of the image content is leveraged within the framework of Markov random fields. Mathematical models are developed around these three topics and constitute building blocks useful for engineering image-based solutions to a wide range of problems.

The contributions include automated quantification of frog patterning from field imagery, statistical methods for estimating the genetic basis of quantified mimicry phenotypes, estimation of the atomic structure of graphene from low-contrast transmission electron microscopy images and patch-based crop classification from synthetic aperture radar data. Further, an information theoretic approach to two-set image decomposition is presented, representing a purely methodological contribution.

This thesis makes statistical image analysis available to fellow researchers with domain specific problems, and provides new methodology relevant for the field itself.

Resumé

Målet for denne afhandling er at udvikle fleksible matematiske metoder til kvantitativ fortolkning af billedindhold. Metodeudviklingen er inspireret af problemer fra så forskellige forskningsområder som evolutionærbiologi, *remote sensing* og materialevidenskab. Løsningerne er inspireret af klassisk matematisk billedanalyse, informationsteori, probabilistiske grafiske modeller og *manifold learning*.

Specifikt omhandler denne afhandling beskrivelse af tre hovedkomponenter i billeder, nemlig intensitet, tekstur og geometri. Modellering af intensitetsfordelinger er vigtig for at opnå nyttige globale repræsentationer af de rå billeddata. Ved hjælp af teksturbeskrivelse kan en lokal repræsentation af billedindholdet opnås, hvilket er nyttigt for deskriptive og diskriminative scenarier. Geometrisk viden om billedindholdet udnyttes gennem Markov random fields formuleringer. Matematiske modeller er udviklet omkring disse tre emner og udgør byggesten, der er nyttige til at konstruere billedbaserede løsninger til et vidt spænd af problemer.

Bidragene inkluderer automatisk kvantifikation af frømønstre fra billeder optaget i felten, statistiske metoder til at estimere den genetiske basis for kvantificerede mimicry-fænotyper, estimering af den atomare struktur af grafen fra transmissions-elektronmikroskopibilleder med lav kontrast, samt patch-baseret afgrødsklassifikation fra syntetisk apertur-radar data. Derudover præsenteres et rent metodisk bidrag i form en informationsteoretisk tilgang til to-sæt billeddekomposition.

Denne afhandling gør statistisk billedanalyse tilgængelig for øvrige forskere med domænespecifikke problemer og bidrager med metodik relevant for feltet i sig selv.

Preface

This thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a PhD in Applied Mathematics with an emphasis on Mathematical Image Analysis. The thesis was prepared with funding solely from the Technical University of Denmark with professor Rasmus Larsen as main supervisor and associate professor Allan Aasbjerg Nielsen as co-supervisor.

The thesis deals with general methods for modelling variability in images, both in the input space and in relevant feature spaces.

The thesis consists of a methodological part, introducing existing state-of-the-art methodology and a part emphasizing the scientific contributions of the thesis.

Lyngby, 31 August2014

Jacob Schack Vestergaard
compute.dtu.dk/∼jsve
github.com/schackv

Papers included in the thesis

[A] J. S. Vestergaard and Allan A. Nielsen. Canonical information analysis, *ISPRS Journal of Photogrammetry and Remote Sensing*, 101:1–9, 2014. doi:10.1016/j.isprsjprs.2014.11.002

[B] J. S. Vestergaard, E. Twomey, R. Larsen, K. Summers and R. Nielsen. Number of genes controlling a quantitative trait in a hybrid zone of the aposematic frog Ranitomeya imitator, in review with *Proceedings of the Royal Society. Series B. Biological Sciences*, 2014.

[C] E. Twomey, J. S. Vestergaard and K. Summers. Reproductive isolation related to mimetic divergence in the poison frog *Ranitomeya imitator*, *Nature Communications*. 5:4749, 2014. doi:10.1038/ncomms5749

[D] J. S. Vestergaard, E. Twomey, A. A. Nielsen, K. Summers, R. Nielsen. Identifying pleiotropic control of adaptive phenotypes, in preparation, 2014.

[E] J. S. Vestergaard, J. Kling, A. B. Dahl, T. W. Hansen, J. B. Wagner and R. Larsen. Structure identification in high-resolution transmission electron microscopy images: an example on graphene, *Microscopy and Microanalysis*, 20(6):1772–1781, 2014. doi:10.1017/S1431927614013464

[F] J. Kling, J. S. Vestergaard, A. B. Dahl, N. Stenger, T. J. Booth, P. Bøggild, R. Larsen, J. B. Wagner and T. W. Hansen. Pattern recognition approach to quantify the atomic structure of graphene, *Carbon*, 74:363–366, 2014. doi:10.1016/j.carbon.2014.03.013

[G] J. S. Vestergaard, A. A. Nielsen, A. L. Dahl and R. Larsen. Classification of Polarimetric SAR Data Using Dictionary Learning, In *Proceedings of SPIE Remote Sensing: Image and Signal Processing for Remote Sensing XVIII*, 8537–35, 2012.
doi:10.1117/12.974814

Other papers by the author

- Mitko Veta, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images, *Journal of Medical Image Analysis*, 20(1):237–248, 2014.
  doi:10.1016/j.media.2014.11.010

- A. B. L. Larsen, J. S. Vestergaard and R. Larsen. HEp-2 cell classification using shape index histograms with donut-shaped spatial pooling, *IEEE Transactions on Medical Imaging*, 33(7):1573–1580, 2014.
  doi:10.1109/TMI.2014.2318434

- J. S. Vestergaard and A. A. Nielsen. Automated Invariant Alignment to Improve Canonical Variates in Image Fusion of Satellite and Weather Radar Data, *Journal of Applied Meteorology & Climatology*, 52(3), 2013.
  doi:10.1175/JAMC-D-12-05.1

- J. S. Vestergaard, A. L. Dahl, P. Holm and R. Larsen. Pipeline for tracking neural progenitor cells, *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, 155–164, 2013.
  doi:10.1007/978-3-642-36620-8_16

- A. A. Nielsen and J. S. Vestergaard. A kernel version of multivariate alteration detection, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3451–3454, 2013.
  doi:10.1109/IGARSS.2013.6723571

- J. S. Vestergaard, A. L. Dahl, P. Holm and R. Larsen. Dynamically constrained pipeline for tracking neural progenitor cells, In *Proceedings of SPIE Medical Imaging: Digital Pathology*, 2013.
  doi:10.1117/12.2006996

- A. A. Nielsen and J. S. Vestergaard. Parameter optimization in the regularized kernel minimum noise fraction transformation, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2012. doi:10.1109/IGARSS.2012.6351561

- A. A. Nielsen, R. Larsen and J. S. Vestergaard. Sparse principal component analysis in hyperspectral change detection, In *Proceedings of SPIE Remote Sensing: Image and Signal Processing for Remote Sensing XVII*, 2011. doi:10.1117/12.897434

Acknowledgements

I would like to thank my supervisors professor Rasmus Larsen and associate professor Allan Aasbjerg Nielsen, with whom I have had the pleasure to work with for more than five years. Their knowledgeable, resourceful and always pleasant guidance has been much appreciated, both on and off campus. Professor Knut Conradsen also deserves thanks for always being supportive and present.

I owe several of my colleagues over the last three years great thanks for their help in preparing this thesis. This includes in particular Stine Harder, Trine Abrahamsen, Anders Nymark, Anders Boesen, Oula Puonti and Line Clemmensen. All of whom I take the liberty of calling good friends.

My collaborators in other research disciplines have been the main motivating force in writing this thesis. Therefore I would like to thank Jens Kling, Jakob Wagner and Thomas W. Hansen at DTU Center for Electron Nanoscopy for directing my attention to the fascinating world at the nanometer scale. An equal thanks to Evan Twomey and Kyle Summers at East Carolina University, without whom I would never have had the opportunity to analyze colorful frogs and dabble in biology.

A special thanks goes out to professor Rasmus Nielsen at University of California Berkeley for hosting me during a very giving research stay in the fall of 2013. It was a great experience, personally as well as professionally. I would also like to mention my lab mates there (Josh Schraiber, Tyler Linderoth, Kelley Harris and others) for being hospitable and inspiring.

Above all I would like to thank my girlfriend Lene for always being patient and tolerating my ever-growing absent-mindedness.

Contents

List of Symbols

$*$	Convolution operator.
$\mathbf{a}, \mathbf{b}$	Linear weightings (projection directions).
$\mathbf{1}_n$	n-vector of ones.
$\mathbf{D}$	A dictionary in $\mathbb{R}^{p \times k}$ of k dictionary atoms.
$\mathbf{I}$	An image in $\mathbb{R}^{m \times n \times q}$.
$\mathbf{I}(\mathbf{x})$	The q-valued observation in the image $\mathbf{I}$ at position $\mathbf{x} = [x, y]$.
$\langle \cdot, \cdot \rangle$	Inner product.
$\mathbf{X}$	Data matrix in $\mathbb{R}^{N \times p}$ with N observations and p variables.
$\mathcal{H}$	Feature space or reproducing kernel Hilbert space (RKHS).
$\mathcal{N}$	Neighborhood system.
$\mathcal{X}$	Input space.
Ω	Parameter space.
$\phi(\cdot, \cdot)$	function mapping from $\mathcal{X}$ to $\mathcal{H}$.
$\mathbf{s}$	Nodes in a graph $\{s_i\}_{i=1}^n$.
σ	Standard deviation or scale parameter.
$\text{corr}(\cdot, \cdot)$	The correlation function.
$\text{cov}(\cdot, \cdot)$	The covariance function.
$\mathbf{x}_0$	Initial start point or vector of starting values.

$i \sim j$	Indicating that sites i and j are neighbors.
$p(\cdot)$	Probability density function.
$q(\cdot)$	Probability density function.
T	Temperature in an annealing scheme.
$U(\cdot)$	Energy function.
$U(\cdot)$	Energy function.
V_C	Potential function.
Z	Partition function.
CCA	Canonical correlation analysis.
cdf	Cumulative distribution function.
CIA	Canonical information analysis.
HRTEM	High-resolution transmission electron microscopy.
ICA	Independent components analysis.
KDE	Kernel density estimator.
LSSHT	Large-scale simultaneous hypothesis testing.
MCMC	Markov Chain Monte Carlo.
MRF	Markov random field.
PCA	Principal components analysis.
pdf	Probability density function.
SAR	Synthetic aperture radar.

Chapter 1

Introduction

Statistical image analysis, as a field, is increasingly relevant for numerous applications. An ever-growing amount of image data are being collected to replace, supplement or improve manual measurements of physical systems. A diverse toolbox of methods is needed to accommodate the analysis of these data in a statistically meaningful way; these methods need to be flexible since it is unfortunately rare that a method developed for one problem, generalize completely to a different problem.

This thesis deals with four cornerstones of image analysis: intensity distributions, extraction of local image information, modelling of geometric structure in images and representation of these in a meaningful feature space. The problems treated in this context are mostly from scenarios where only limited reference data are available, no more data will become available, and manual labelling is extremely tedious or even impossible. In such cases, a mathematical model needs to enforce the known constraints and use the image information in the light of this prior knowledge. In other words, all available information should be used to solve the problem.

1.1 Motivating examples

The work in this thesis is largely motivated by collaborations. Especially collaborations with fellow researchers in other disciplines trying to answer questions in their respective domains. Often they lack the ability to quantitatively interpret the images they have at hand. These problems have spurred the development of the statistically sound methodology for image analysis around which this thesis revolves. Some motivating examples will be presented below.

1.1.1 Intensity distributions

The distribution of image intensities is a statistical description of the contents of an image. It does not readily model any spatial properties, but merely summarizes the numeric content. This makes it a fundamental tool for checking assumptions about the data, e.g., are they normally distributed, are they bimodal, etc.

Different image modalities exhibit different distributions, which always needs to be considered and in some cases can be exploited. An example of this can be seen in Figure 1.1, where the distributions of weather radar reflectances and three satellite (near) infra-red bands over the same geographical region and acquired at the same time are shown. Clearly the distributions differ. This is due to the different aspects that are captured by these modalities. These differences need to be considered when comparing or combining data from different modalities. For instance, it does not make sense to do direct numerical comparisons between the distributions in Figure 1.1 even though the underlying physical phenomenon is the same.

Summary statistics of distributions can be very useful for interpretation and numerical optimization. For instance, it is a lot less tedious to talk about a normal distribution in terms of its mean $\mu = 1$ and variance $\sigma^2 = 1.5$, than in terms of the full distribution. For optimization purposes, the optimality condition is often given in terms of the extremum of a suitable summary statistic, e.g., variance for the purpose of principal components analysis. To uncover associations between different distributions, measures such as correlation and mutual information will prove useful. This is the type of problem around which Paper A revolves.

Describing images in terms of their intensity distributions are very relevant for the purpose of image decomposition, which will be motivated in Section 1.1.4.

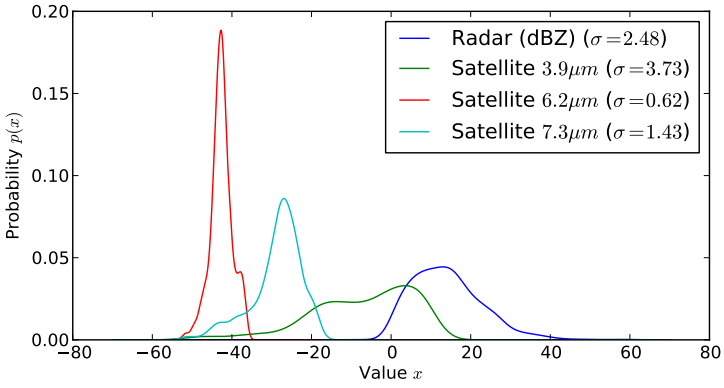


Figure 1.1: Distributions of weather radar reflectances (measured in dBZ) and weather satellite brightness temperatures for three (near) infrared bands. The shown distributions are kernel density estimates using the maximal smoothing principle for bandwidth estimation (see Section 2.1.2). The images underlying these distributions are shown in Figure 1.5.

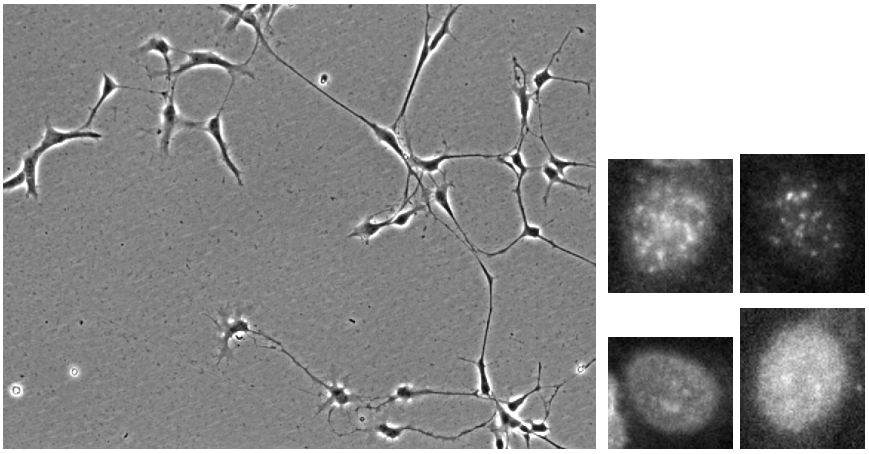
1.1.2 Capturing texture variability using descriptors

The intensity distribution described above is a global description, i.e., it summarizes the entirety of the signal. Often it is desirable to describe local information, which can be achieved by individual distributions of spatially confined regions. This is the underlying principle of many image descriptors.

The relevant information to extract from an image depends on the purpose of the analysis. Thus in many cases, the ability to capture different aspects of the variability in images becomes necessary. Images are at least two-dimensional and, depending on the image formation process, the scale of the relevant objects in the image varies. Further, images are inherently large-scale, since even a small image of size 256×256 pixels yields 65,536 observations. These are all conditions that the image data extraction needs to take into account.

Several examples can be given of the relevance of capturing image variability with local texture descriptors. For instance, detecting neural progenitor cells in phase contrast microscopy images (Figure 1.2) by segmenting the image into cell/not-cell areas can be a difficult task, when the contrast is low. In that case it might be useful to use an entire image patch as the local description of the image, rather than a single pixel, to encode spatial context. This is used in Vestergaard et al. (2013a) and Vestergaard et al. (2013b). A similar example is

the motivation for Paper G, where the task is to classify a polarimetric synthetic aperture radar (SAR) image into different types of crops.



(a) Neural progenitor cells. **(b)** HEp-2 cells.

Figure 1.2: Examples of different types of interesting image information. For neural progenitor cells, the cell somas are of main interest, while for HEp-2 cells the texture signature is more important.

The problem of classifying different types of cells yields a major problem in histopathology, since a manual annotation is tedious and subject to large inter- and intra-observer variances. This motivates the need for automatic classification methods. It is important that such classification methods describe the relevant variability, where it is common to extract a surplus of image features and use a training/validation data scheme to tune the importance weighting of each feature. As such it is useful if the image features separate well, i.e., they describe a controllable limited aspect of the image. The papers Larsen et al. (2014) and Veta et al. (2014) leverage this to achieve high performance classification.

A different motivation is from biology, where the aposematic frog *Ranitomeya imitator* in one end of a transect mimics one dendrobatid species, while in the other end it mimics another. The mimicry is a color pattern polymorphism, which needs to be quantified objectively. In this case the images are JPEG compressed field photographs, the lighting is sub-optimal, and the frogs are in different poses. Under these circumstances, how is the pattern and color best quantified as a basis for answering important evolutionary hypotheses? See Figure 1.3 for some examples.

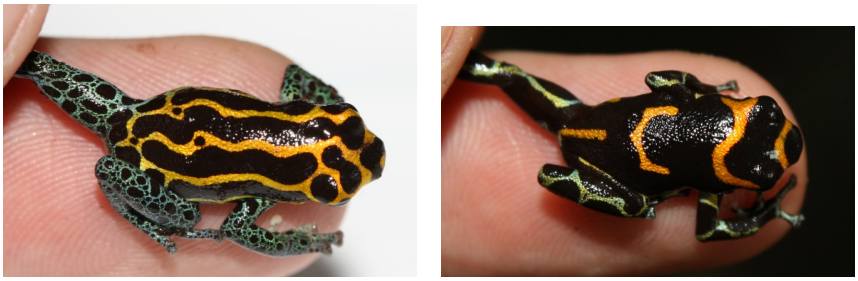


Figure 1.3: For these colorful poison dart frogs of the species *Ranitomey imitator*, the nature of the pattern, e.g., the directionality of the stripes, is interesting to quantify.

1.1.3 Inferring geometry

Recent years' reduction of cost for high-resolution microscopes for research purposes has brought an increased need for image analysis in materials science. An interesting problem is that of identifying the microscopic structure of the up-and-coming material graphene from high-resolution transmission electron microscopic (HRTEM) images. The interest in the microscopic structure is spurred by the coupling between the microscopic structure and the macroscopic properties, such as conductivity and strength. Due to the image formation process, the hexagonal structure of the carbon atoms are not directly visible, but need to be inferred from what little contrast is present. Luckily, there is a strong prior knowledge about how atoms can form in such lattices. Examples of such a HRTEM image can be seen in Figure 1.4.

Incorporating this knowledge in an applicable algorithm was the underlying motivation of the initial work in Kling et al. (2013) and the solution proposed in Papers E and F.

1.1.4 Descriptive subspaces

Having quantified the image variability important for a given problem, what is the best space to represent this data? The answer to this question depends on the hypothesis of interest. E.g., for visualization purposes, the best embedding might be in a subspace defined by the first few principal components, while for discriminative purposes, the best embedding might be one maximizing class separability. Learning an appropriate subspace, manifold or embedding of a set of features is motivated by and used in multiple cases.

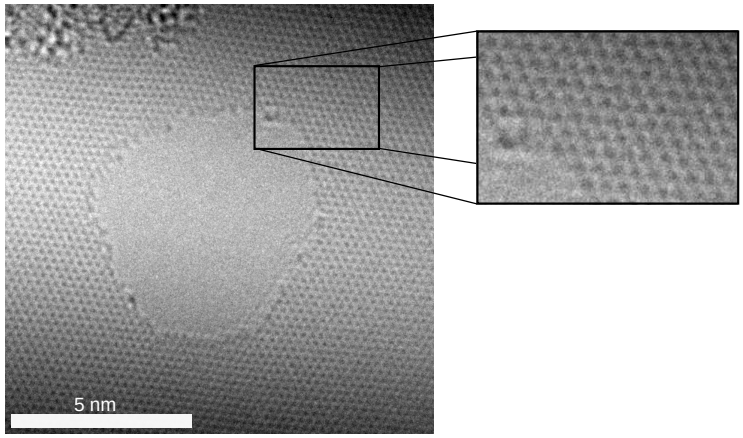
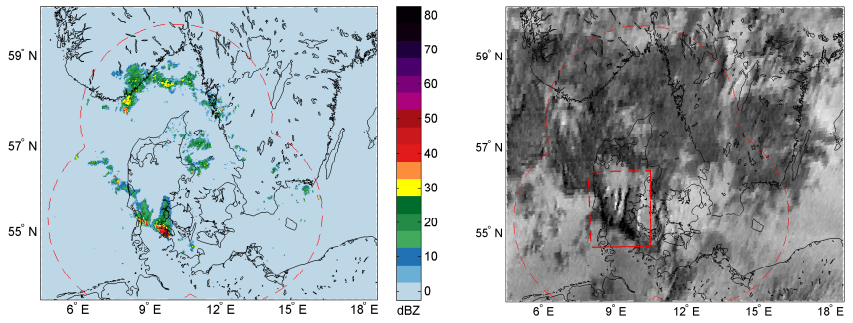


Figure 1.4: High-resolution transmission electron microscopy image of a graphene sample with an induced hole. A region is magnified to better see the contrast properties in the image. The dark spots are not carbon atoms, but rather the space in the middle of the carbon hexagons.

A first example is detection of extreme rain from satellite imagery. In the paper Vestergaard and Nielsen (2012) we found that canonical correlation analysis (CCA) could successfully find a linear decomposition of the original eight infrared bands of satellite imagery to enhance the contrast between extreme rain and uninteresting cloud coverage in a one-dimensional embedding. However, it required a very elaborate geometrical and temporal alignment of the radar and satellite imagery prior to this decomposition. See Figure 1.5 for an example of the non-informative CCA solution without the geometrical alignment, which spurred the need for a different approach. The motivation for Paper A is to leverage the information theoretical concepts of entropy and mutual information to provide a decomposition taking the distribution of the two very different image modalities into account. This is useful not only for these two modalities in particular, but decomposition of any two (non-Gaussian) sets of variables can benefit from such a method.

Returning to the example with the aposematic frog *Ranitomeya imitator* from above. Having quantified the color and pattern of these frogs, it must be determined to what extent the different aspects of this phenotype relates to mimicry. This is an example of having a surplus of image features extracted to avoid inducing subjective biases (e.g., only measure what is thought to be important) and thus having an overcomplete representation. The reduction to a low-dimensional manifold must be carried out in conjunction with the biological problem statement. This is of relevance to Papers B, C and D.



(a) Radar data. **(b)** Satellite data (CCA).

Figure 1.5: Two image modalities covering the same area at the same time. (a) The single weather radar data band showing reflectances. (b) Linear combination of the eight infra-red bands of satellite data that maximizes correlation with the radar data, found using canonical correlation analysis (CCA). The marked rectangular area would ideally exhibit high contrast between rain and no-rain areas.

1.2 Thesis objectives

The main objectives of this thesis are:

- Develop or improve existing state-of-the-art methodology in applied image analysis.
- Engineer flexible image-based solutions to cross-disciplinary research problems.
- Investigate the generalizability and flexibility of methods to model image variability in the input space, in terms of intensity, texture and geometry. And choosing appropriate feature spaces for given design criteria.

These objectives have been sought accomplished mainly through collaborations on a selection of interesting problems in the life sciences and materials science. Isolated methodological development has also been performed, in which case applicability has been illustrated through use of relevant examples.

1.3 Reading guidelines

The present thesis consists of two parts: A part on methodology and a part on scientific contributions employing and extending this methodology. Conclusions on the thesis are provided, before the included papers are appended.

Methodology The methodology part (Part I) is not self-contained. It contains the prerequisites for understanding the context of the scientific contributions. As such, the scientific contributions contain new and extensions of existing methodology useful for solving specific problems in various domains. Appendices are included with methodology that is relevant, but not a major part of the thesis.

Scientific contributions Part II summarizes the scientific contributions by topic. This means that several papers are summarized together, when they share both methodology and application. The summaries provide a general motivation for solving the treated problems and give an overview of the solutions presented in the papers. The main results and contributions are highlighted. Papers included as part of the thesis are appended as Papers A–G.

Reading flow The intention with the thesis structure is that the reader can either read from the beginning to the end, or read a summary of a scientific contribution of interest and refer to Part I and the relevant papers for detail.

Notation While most notation should be clear from context, a nomenclature is provided in the preface, which should aid in interpreting notation and abbreviations that are less obvious. Note that common biological notation has been adopted such that genus names will be abbreviated on subsequent mentions.

Code I provide publicly available MATLAB or Python 3.x implementations of the methods used, for most of the published papers, at `https://github.com/schackv`.

Part I

Methodology

Introduction to methodology

Observation of physical phenomena by image acquisition is often a sensible choice, since the data can be stored and analyzed later. However, images can often be challenging to analyze. Data volume, imaging device, scale, contrast, spatial context and the limited amount of manually annotated data are all properties that must be taken into consideration. One single method for analysis of the image data is not enough; rather a multitude of building blocks for image processing pipelines are necessary to successfully tackle real-world image analysis problems. The methodology presented in Chapter 2 considers how three major aspects of image data can be modelled, namely intensity, texture and geometry. In Chapter 3 the focus will be on methods for learning an appropriate feature space for investigating a given hypothesis.

Chapter 2

Intensity, texture and geometry

This chapter describes some fundamental components of image analysis, in the image input space, namely intensity, texture and geometry.

Intensities, i.e., the pixel values, are the basic data entity in images. Commonly, images come in either gray-scale or RGB, providing one or three channels as a representation of a scene, while in fields such as remote sensing, it is common to analyze imagery from multi- or hyperspectral sensors. The statistical distribution of the intensities often needs to be modelled, either to perform hypothesis testing, do visualization or to choose a statistically meaningful subspace for further analysis.

Texture is a common term for the distribution of intensities in a spatially confined neighborhood of an image. Rather than attempting to model the particular instance of the region, the texture of an image region says something about the nature of the variation. Texture may vary in an image, depending on the scale in which it is observed. Image descriptors and image patches are two of the common ways of extracting texture information from images.

When fitting a spatial model to an image, the geometry of this model often needs to be constrained, i.e., the prior information (or assumption) about the

model needs to be leveraged. This model could be a binary segmentation of a gray-scale image, fitting a shape model of a face to a portrait photo or fitting a grid structure to an observed instance. In all cases, models are more or less based on our prior knowledge: pixels of value 0 should probably be close to each other; the nose should probably be between the mouth and the eyes, and the grid structure should not fold on top of itself. These spatial constraints can be modelled in the framework of Markov random fields, which will be described in Section 2.3.

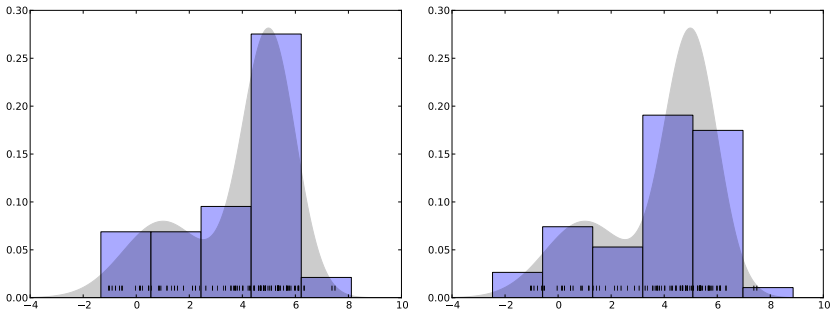
2.1 Image intensity distributions

Different image modalities exhibit different distributions of intensities. For instance, optical images can often be assumed to be normally distributed in each band. In medical image modalities such as magnetic resonance imaging (MRI), computed tomography (CT) or positron emitting tomography (PET) the distributions are known to be non-normal (Rician (Gudbjartsson and Patz, 1995) and Poisson (Buzug, 2008) respectively). Various Earth observation modalities are also known to be inherently non-normal in nature: for instance LiDAR that measures distances, weather radar data measuring reflectance properties or polarimetric synthetic aperture radar (SAR), where complex covariance matrices in each pixel follows a complex Wishart distribution (Conradsen et al., 2003).

2.1.1 Histogram

The histogram is the most common representation of a distribution of intensities. It is a discrete representation of samples $\{x_i\}_{i=1}^{N}$ from a continuous or discrete distribution. It consists of B bins with heights $\{b_j\}_{j=1}^{B}$ and edges $\{e_k\}_{k=1}^{B+1}$. The height of each bin is the number of samples where $e_i \leq x_i < e_{i+1}$. Normalizing such that $\sum_{j=1}^{B} b_j = 1$ ensures that the histogram can be seen as a probability density function (pdf).

The domain and the number of bins must be chosen when fitting the histogram, either manually or by some heuristic, e.g. Scott's rule (Scott, 1979). Too few bins might hide important aspects of the distribution, while too many bins will overfit to the sample rather than represent the underlying distribution. An example of how subtle changes in binning can change the appearance of the histogram, and thus the estimate of the pdf, drastically can be seen in Figure 2.1.



(a) Histogram. **(b)** Histogram, bins shifted.

Figure 2.1: Impact of binning choice differences. The $N = 100$ samples are shown as ticks and the true underlying distribution is the filled gray area.

The histogram can approximate a continuous pdf, since for B bins of equal width Δ covering the domain of X

$$\lim_{\Delta \to 0} \sum_{j=1}^{B} b_j \Delta \to \int_{\Omega} p(x) \mathrm{d}x \ . \tag{2.1}$$

However, only a finite number of samples from the true pdf is available, wherefore a non-infinitisemal bin width is necessary, thus rendering the histogram an approximation only. The normalization of a histogram representing a continuous distribution factors in the bin width such that $\sum_{j=1} b_j \Delta = 1$ to ensure that $\int p(x) \mathrm{d}x = 1$ in the limit (Bishop, 2007).

2.1.2 Kernel density estimation

The histogram is a simple non-parametric density estimator. However, as seen above the estimated histogram is not smooth and it depends on the bin end points and width. By using kernel density estimators (KDE) (Rosenblatt, 1956, Parzen, 1962, Silverman, 1986) where we center a kernel on each observation, we may obtain smoother histograms that do not depend on bin end points. The kernel density estimator (Parzen windows estimator) for the pdf of X at value t is

$$\hat{p}(X = t|\mathbf{x}) = \frac{1}{N\sigma} \sum_{i=1}^{N} \varphi \left(\frac{t - x_i}{\sigma} \right) \tag{2.2}$$

where $\varphi(z)$ is the kernel and σ is a smoothing parameter referred to as the bandwidth. The $\hat{}$ indicates that this is an estimator of the distribution. Often

the Gaussian kernel

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{2.3}$$

is chosen. The width of the Gaussian, i.e., the standard deviation is thus equivalent to the bandwidth σ.

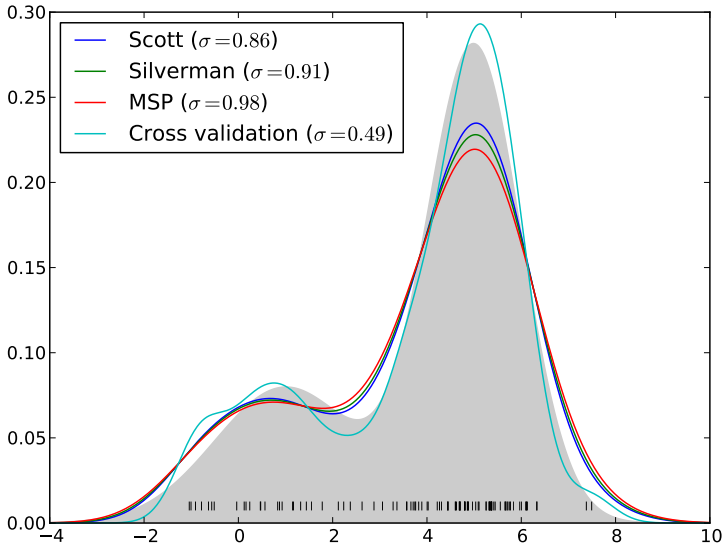


Figure 2.2: Kernel density estimates of distribution with different bandwidth estimators with the bandwidth σ indicated. The $N = 100$ samples are shown as ticks and the true underlying distribution filled in gray.

Estimation of the bandwidth is an example of the bias-variance trade-off: a too narrow kernel causes too large variation in the density estimate and a too wide kernel oversmooths the estimated distribution (Jones and Marron, 1996). In Figure 2.2 some of the well-known bandwidth estimators are used to fit a KDE with a Gaussian kernel to the same data as in Figure 2.1. The Scott and Silverman estimators are so-called rule-of-thumb estimators (Scott, 1979, Silverman, 1986), while the maximal smoothing principle (MSP) is based on the interquartile range of the sample (Terrell, 1990). The cross validation approach exploits the fact that given a set of parameters, the probability of observing a set of points can be interpreted as a likelihood of the parameters. Here a 20 fold cross validation scheme is used, where one twentieth of the samples are left out successively and used to get likelihood estimates of 30 equidistantly spaced bandwidths in the range of $[0.01, 2]$.

Botev et al. (2010) point out certain shortcomings with several of the existing

kernel density bandwidth estimators. For instance, the popular Sheather-Jones plug-in estimator (Sheather and Jones, 1991) relies on an inital assumption of Gaussianity, which is conceptually unsatisfactory even though it often works well in practice. Further, the Gaussian kernel density estimator lacks local adaptivity, due to its smoothing properties. Other estimators do not yield genuine pdfs, either due to violating the non-negativity constraints or not integrating to one. Botev et al. (2010) present a KDE based on linear diffusion processes, which shows promising performance, but was found too sensitive to small changes for the work presented in Paper A.

2.1.3 Statistics

Various statistics can be used to summarize a distribution from its samples $\{x_i\}_{i=1}^N$ or measure the association between two random variables. For instance, the sample variance

$$\hat{\sigma}^2(X) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \quad \text{where } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.4}$$

is an unbiased estimator of the population variance.

The variance is particularly interesting for normally distributed data and describes completely the spread of such data, while other statistics are relevant to describe, e.g., the degree of information content of a distribution. Spatial properties of an image are also interesting, which will be treated in Section 2.2.

Covariance and correlation are two useful measures of association between two variables X and Y. With N observations from the two variables, the sample covariance is

$$\widehat{\mathrm{cov}}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{2.5}$$

or $\frac{1}{N-1}\mathbf{x}^T\mathbf{y}$ for zero-mean vectors. Correlation is the covariance normalized by the spread of the two variables

$$\widehat{\mathrm{corr}}(X,Y) = \frac{\widehat{\mathrm{cov}}(X,Y)}{\sqrt{\widehat{\sigma^2}(X)\widehat{\sigma^2}(Y)}}\,. \tag{2.6}$$

2.1.3.1 Entropy

Entropy was introduced in information theory for discrete variables by Shannon (1948) as a measure of a signal's information content and thus the needed bits to encode the information: a high entropy meaning high information content and thus more bits needed, and a low entropy meaning low information content and thus less bits needed. We will here distinguish between discrete entropy $H(X)$ and differential (or continuous) entropy $h(X)$ and start by introducing discrete entropy.

Definition 2.1 (Discrete entropy). Discrete marginal entropy of a discrete random variable X is defined from the expectation expectation $\mathbb{E}[-\ln(P(X = x))]$ as

$$H(X) = -\sum_x P(X = x) \ln P(X = x) \, . \tag{2.7}$$

Discrete joint entropy of two discrete random variables X and Y is defined as

$$H(X, Y) = -\sum_x \sum_y P(X = x, Y = y) \ln P(X = x, Y = y) \, . \tag{2.8}$$

Properties of discrete entropy include:

$$\begin{aligned} H(X) &\geq 0 \\ H(X, Y) &\geq \max(H(X), H(Y)) \geq 0 \\ H(X, X) &= H(X) \\ H(X, Y) &\leq H(X) + H(Y) \quad \text{(Equal iff } X \text{ and } Y \text{ are independent)} \, . \end{aligned}$$

The unit of discrete entropy dependes on the logarithm used, e.g., bits when using $\log_2$ and nats when using the natural logarithm. ▲

Entropy as a measure of the distribution of a continuous random variable is termed *differential entropy*. Differential entropy is defined in Definition 2.2. As opposed to discrete entropy, differential entropy is not necessarily non-negative since it is possible that $p(x) > 1$.

Definition 2.2 (Differential entropy). Differential marginal entropy of a continuous random variable $X : \Omega \mapsto \mathbb{R}$ with pdf $p(x)$ is defined from the expectation $\mathbb{E}[-\ln(p(x))]$ and can be estimated from the pdf as

$$h(X) = -\int_\Omega p(x) \ln p(x) \mathrm{d}x \, , \tag{2.9}$$

or as a sample estimate $\hat{h}(X) = -\frac{1}{N} \sum_{i=1}^N \ln(p(x_i))$ from the N samples $\{x_i\}_{i=1}^N$. Differential joint entropy of two discrete random variables $X : \Omega_X \mapsto \mathbb{R}$ and

$Y : \Omega_Y \mapsto \mathbb{R}$ is defined as

$$h(X,Y) = -\int_{\Omega_X}\int_{\Omega_Y} p(x,y)\ln p(x,y)\mathrm{d}y\mathrm{d}x\ . \tag{2.10}$$

Properties of differential entropy include:

$$h(X,X) = h(X)$$
$$h(X,Y) \leq h(X) + h(Y) \quad \text{(Equal iff } X \text{ and } Y \text{ are independent)}\ .$$

▲

The distribution with maximum entropy is thus the uniform distribution, since every value within the domain is equally likely to appear. In relation to information theory, this also means that the least compression can be achieved. Specifically the differential entropy of a random variable X distributed evenly on the interval $[a,b]$, i.e., with density $p(x) = \frac{1}{b-a}$ is

$$h(X) = \int_a^b \frac{1}{b-a}\ln\frac{1}{b-a}\mathrm{d}x = \ln(b-a)\ . \tag{2.11}$$

This also illustrates that entropy can be negative: if $b-a < 1$ then $h(X) < 0$. For a given mean and variance, the normal distribution is the continuous distribution with maximum entropy (Bishop, 2007, Cover and Thomas, 2006).

The differential entropy can be estimated using a kernel density estimator $p_\varphi(x)$ with kernel $\varphi(z)$ and scale σ for the pdf such that

$$h_\varphi(X) = -\sum_{i=1}^{N} p_\varphi(x_i)\ln p_\varphi(x_i)$$
$$p_\varphi(x_i) = \frac{1}{N\sigma}\sum_{i=1}^{N}\varphi\left(\frac{x_j - x_i}{\sigma}\right)\ .$$

Differential entropy is not the limit of quantized entropy A peculiar relation exists between differential and discrete entropy, namely that the discrete entropy of a pdf divided into infinitesimally small bins does not yield the differential entropy in the limit. This result is also stated and described by Cover and Thomas (2006).

Consider a continuous pdf $p(x)$ quantized in N bins of width Δ. The mean value theorem tells us that there exists an x_i such that

$$p(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} p(x)\mathrm{d}x\ . \tag{2.12}$$

Thus the probability that X falls in the ith bin is $p_i = p(x)\Delta$. The discrete entropy of this quantized pdf is

$$H_\Delta(X) = -\sum_{i=1}^{N} p_(x_i)\Delta \ln(p_(x_i)\Delta) = -\sum_{i=1}^{N} p_\Delta(x_i)\Delta \ln p_\Delta(x_i) - \ln\Delta \quad (2.13)$$

since $\sum_{i=1}^{N} p(x_i)\Delta = 1$. Consider now the discretization of differential entropy in bins of width Δ

$$-\int_\Omega p(x)\ln p(x)\mathrm{d}x \approx -\sum_{i=1}^{N} \Delta p(x)\ln p(x)\ . \quad (2.14)$$

We see how this differs from Eq. (2.13) by $-\ln\Delta$. In general this shows that

$$\lim_{\Delta\to 0} H_\Delta(X) \to h(X) - \ln\Delta\ . \quad (2.15)$$

The same result can be derived for joint entropy in which case

$$\lim_{\Delta_X,\Delta_Y\to 0} H_{\Delta_X,\Delta_Y}(X,Y) \to h(X,Y) - \ln\Delta_X\Delta_Y\ . \quad (2.16)$$

where Δ_X and Δ_Y are bin widths in each direction.

2.1.3.2 Kullback-Leibler divergence and mutual information

The *Kullback-Leibler divergence* (or *relative entropy*)

$$D_{\mathrm{KL}}(p||q) = \int p(x)\ln\frac{p(x)}{q(x)}\mathrm{d}x \quad (2.17)$$

is a measure of association between two pdfs $p(x)$ and $q(x)$. D_{KL} is always non-negative and zero only if p and q are independent. Due to lack of symmetry ($D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$) this is not a true distance measure (Cover and Thomas, 2006).

Mutual information (MI) describes the association between two random variables X and Y in terms of "shared information". MI is defined as the relative entropy between the joint distribution $p(x,y)$ and product of marginals $p(x)p(y)$ such that

$$I(X,Y) = \int_x\int_y p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\mathrm{d}y\mathrm{d}x\ . \quad (2.18)$$

The motivation is that if the product of the marginals completely describes the joint distribution then the two distributions are independent and the mutual

information is zero. This expression can be expanded to

$$\begin{aligned}
I(X,Y) &= \int_x \int_y p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \mathrm{d}y\mathrm{d}x \\
&= \int_x \int_y p(x,y) \left(\ln p(x,y) - \ln p(x) - \ln p(y)\right) \mathrm{d}y\mathrm{d}x \\
&= \int_x \int_y p(x,y) \ln p(x,y) \mathrm{d}y\mathrm{d}x - \int_x \int_y p(x,y) \ln p(x) \mathrm{d}y\mathrm{d}x \\
&\quad - \int_x \int_y p(x,y) \ln p(y) \mathrm{d}y\mathrm{d}x \\
&= - \int_x p(x) \ln p(x) \mathrm{d}x - \int_y p(y) \ln p(y) \mathrm{d}y \\
&\quad + \int_x \int_y p(x,y) \ln p(x,y) \mathrm{d}y\mathrm{d}x \\
&= h(X) + h(Y) - h(X,Y)
\end{aligned} \tag{2.19}$$

which gives us the preferred form of mutual information. A variety of normalizations have been proposed (see e.g., Astola, 1982, Yao, 2003, Witten and Frank, 2005, Vinh et al., 2010); some examples are

$$I^n(X,Y) = 2\frac{I(X,Y)}{h(X) + h(Y)} \tag{2.20}$$

$$I^n(X,Y) = \frac{I(X,Y)}{h(X,Y)} \tag{2.21}$$

$$I^n(X,Y) = \frac{I(X,Y)}{\sqrt{h(X)h(Y)}} \,. \tag{2.22}$$

Care should be taken when normalizing mutual information based on quantized entropy estimates of the differential entropy. While substituting the differential entropy for a quantized estimate (Eqs. (2.15) and (2.16)) into Eq. (2.19) yields

$$\begin{aligned}
I_\Delta(X,Y) &= h(X) - \ln \Delta_X + h(Y) - \ln \Delta_Y - h(X,Y) + \ln \Delta_X \Delta_Y \\
&= h(X) + h(Y) - h(X,Y) = I(X,Y)
\end{aligned}$$

where it is seen that the discrepancies cancel out, i.e., the quantized entropy can be used as direct substitute for the differential entropy when calculating mutual information. However, plugging into for instance the normalization in Eq. (2.21) yields

$$I^n_\Delta(X,Y) = \frac{I(X,Y)}{h(X,Y) - \ln \Delta_X \Delta_Y}$$

and is seen to provide a different normalization than expected.

2.2 Image texture descriptors

Image descriptors is a common name for a class of local image models. Descriptors are typically designed to extract some sort of feature at or around a given point. This feature is usually very basic and could be shape, color or texture related. Feature extraction is useful in diverse scenarios, such as object recognition (Lowe, 1999, Mikolajczyk and Schmid, 2005, Bay et al., 2006, Tola et al., 2010), structure from motion (Harris, 1993, Lowe, 2004), where features are extracted in successive images and point correspondances are determined to infer the three dimensional structure, and in (supervised) predictive contexts (Lin et al., 2011, Larsen et al., 2014, Veta et al., 2014), where different descriptors are used to capture different aspects of the observed image and their importance for the prediction is learned from training data. Larsen (2012) provides a neat introduction to local image descriptors and Mikolajczyk and Schmid (2005) provides a performance comparison of local descriptors in the context of object and scene recognition.

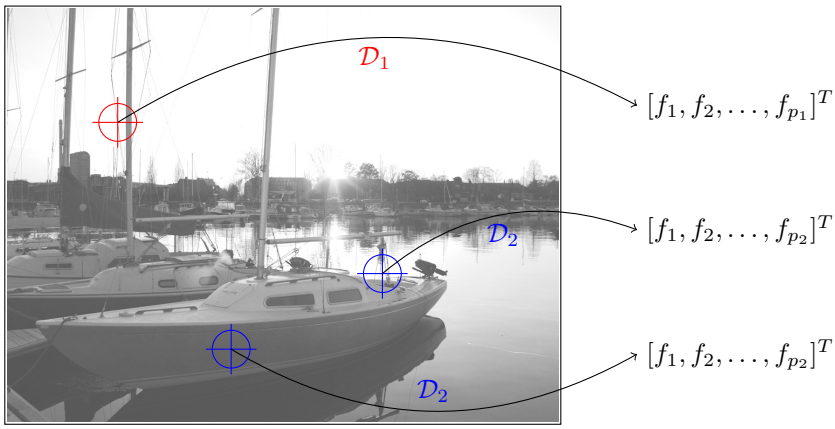


Figure 2.3: Extraction of image features using two descriptors $\mathcal{D}_1$ and $\mathcal{D}_2$, each giving features of dimensionality p_1 and p_2 respectively.

In general, image descriptors are tools for capturing image variability, i.e., going from the image domain to a feature space. This is sketched in Figure 2.3. An image descriptor $\mathcal{D}$ maps from a point $\mathbf{x} = (x, y)$ in the (q-variate) image $\mathbf{I} \in \mathbb{R}^{m \times n \times q}$ to a feature space and can be univariate $\mathcal{D} : \mathbf{I}(\mathbf{x}) \mapsto \mathbb{R}$ or multivariate $\mathcal{D} : \mathbf{I}(\mathbf{x}) \mapsto \mathbb{R}^p$. When extracting multiple descriptors these are simply concatenated and thus extending the dimensionality of the feature space . In descriptive and predictive contexts a single descriptor will be a too limited model of the variability, while too many descriptors increase the dimensionality p of the feature space and thus increase the risk of overfitting, i.e., in the choice

of descriptor suite lies an example of the bias-variance trade-off (Hastie et al., 2003). In a supervised scenario with plenty of data (large N) insignificant descriptors can easily be learned to weigh less on predictions, while unsupervised descriptive scenarios are inherently harder to handle.

Common for most image descriptors is that they are extracted at some scale of the original image. Here we briefly review gradient orientation histograms and shape index histograms in a scale-space setting (Lindeberg, 1996) with notation fom the locally orderless images (LOI) (Koenderink and Doorn, 1999) framework.

2.2.1 Scale-space and locally orderless images

Lindeberg (1993) motivates the need for a scale-space nicely:

> Why should one represent a signal at multiple scales when all information is anyway in the original data? A major reason for this is to explicitly represent the multi-scale aspect of real-world data. Another aim is to suppress and remove unnecessary and disturbing details, such that later stage processing tasks can be simplified.

The scale-space representation of the image $\mathbf{I}(\mathbf{x})$ is defined as

$$L(\mathbf{x}, \sigma) = (G * \mathbf{I})(\mathbf{x}; \sigma) \tag{2.23}$$

where $\mathbf{x} = [x, y]$ and σ is the scale. $(G * \mathbf{I})$ is a convolution of the image with a Gaussian kernel with standard deviation σ. The Gaussian kernel is unique for generating a scale space (Koenderink, 1984, Lindeberg, 1996) since it is the solution to the diffusion equation

$$\partial_\sigma = \frac{1}{2}(\partial_{xx} + \partial_{yy})L \quad \text{s.t. } L(\cdot; 0) = \mathbf{I}\,.$$

This also implies that the original image can be seen as an initial heat distribution evolving over time (scales) in homogeneous medium, thus providing a natural dissolution of fine details. Examples of scale-space representations can be seen in Figure 2.4

The image derivatives can be calculated in this scale space formulation, where L_x and L_y denote the gradient in the x- and y-direction respectively and the Hessian matrix

$$\nabla^2 L = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{bmatrix} \tag{2.24}$$

(a) $\sigma = 2.5$ **(b)** $\sigma = 10$

Figure 2.4: Scale-space representations of a 768×1024 example image.

describes the local curvature. We omit $\mathbf{x}$ from the left hand side of the definitions below for brevity.

The framework of locally orderless images (LOI) was introduced by Koenderink and Doorn (1999) in which the spatial order of pixels within a region-of-interest (ROI) around a probed point is disregarded and the texture reduced to a form summarizable by histograms. Three notions of scale are in play when working with LOI: the *inner scale*, *outer scale* and *tonal range*. The scale space introduced above lets us probe the image at different inner scales (or resolutions). The outer scale (or *aperture width*) α defines the ROI by weighing contributions from nearby pixels with their distance to the probed point; the weighting is analogous to a Parzen window density estimate introduced in Section 2.1.2. The tonal range β defines the third relevant scale, namely a soft weighting similar to a bin width; say the histogram captures the intensity of pixels, then a small β would resemble narrow bins and thus a lot of bins would be needed. Here the LOI framework is used to capture local characteristics of image first and second order derivatives.

2.2.2 Gradient orientations and the shape index

Gradient orientations have been widely popular in computer vision to capture local geometry (Lowe, 1999, Dalal and Triggs, 2005, Mikolajczyk and Schmid, 2005, Avidan, 2006, Bosch et al., 2007). Part of the popularity is due to the invariance of local derivatives towards global illumination changes.

Gradient magnitude m and orientation θ can be derived from L_x and L_y as

$$m = \sqrt{L_x^2 + L_y^2}\ , \quad \theta = \text{atan2}(L_x, L_y)\ . \tag{2.25}$$

The gradient orientation is circular on the interval $]-\pi, \pi]$. To quantify the amount of first order change in a given orientation, the gradients are quantized in q bins centered at $b_i, i = 1, \ldots, q$ in this interval.

At a given scale σ, for bin b the gradient orientation descriptor is defined as

$$\text{goh}(b;\sigma) = m(\mathbf{x};\sigma)\frac{\exp\left(\beta^{-2}\cos\left(\theta(\mathbf{x};\sigma) - b\right)\right)}{2\pi I_0(\beta^{-2})} \tag{2.26}$$

Due to the cyclic nature of the gradient orientations the von Mises aperture is used, where $I_0()$ is the modified Bessel function of order 0 and β is the tonal range. See Koenderink and Doorn (1999), Larsen et al. (2014) for more details. Note that the gradient orientation contribution is weighted by its magnitude m, which ensures that well defined gradients count more than spurious ones.

2.2.2.1 Shape index

The shape index is a second order image descriptor used to describe local curvature (Koenderink and van Doorn, 1992) and is derived from an eigendecomposition of the local Hessian. The eigenvectors capture the orientation of the curvature, while the eigenvalues capture the nature of the curvature (cup, cap, saddle, etc.). For the shape index definition, the orientation is omitted and thus enforcing rotation invariance.

The shape index s is defined as

$$s = \frac{2}{\pi}\text{atan}\left(\frac{-L_{xx} - L_{yy}}{\sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}}\right)\ , \quad s \in [-1, 1] \tag{2.27}$$

with curvature $c \in \mathbb{R}_+$

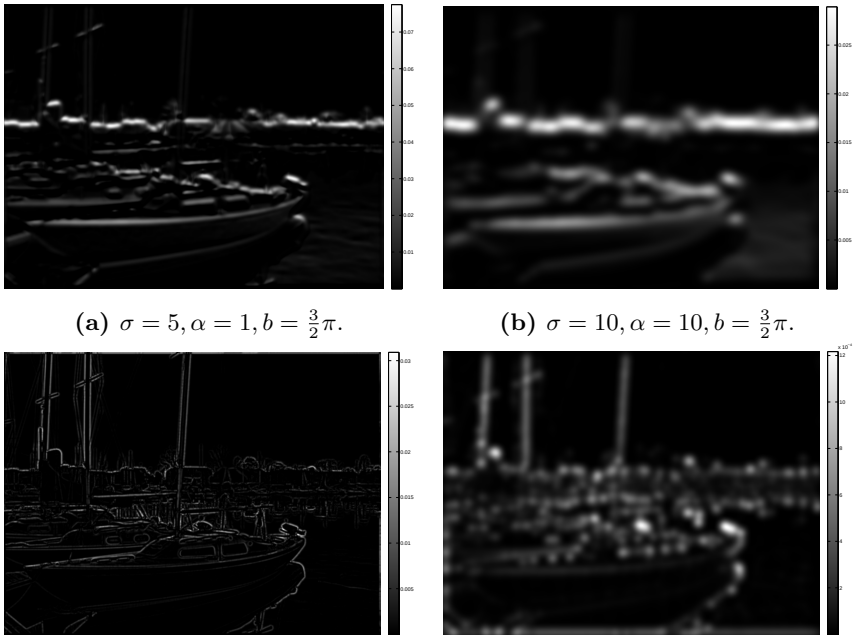
$$c = \frac{1}{2}\sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}\ . \tag{2.28}$$

The binning of the shape index into a histogram is similar to that of the gradient orientation histograms. However, the range of the shape index is not cyclic

wherefore a standard Gaussian aperture function can be used. The shape index histogram descriptor for bin b at scale σ is

$$\mathrm{sih}(b;\sigma) = \frac{c(\mathbf{x})}{2\pi\beta^2} \exp\left(-\frac{(s(\mathbf{x})-b)^2}{2\beta^2}\right) . \tag{2.29}$$

Examples of the gradient orientation and shape index reponse for a few different parameters can seen in Figure 2.5. The gradient orientation can be seen to highlight horizontal lines in the image, since the chosen bin $b = \frac{3}{2}\pi$ is perpendicular to this. The shape index is seen to capture both ridge-like curvatures for bins at $b = 0.4$ and cup-like responses for $b = -0.8$. The scale space dependence is apparent from these examples.



(a) $\sigma = 5, \alpha = 1, b = \frac{3}{2}\pi$.

(b) $\sigma = 10, \alpha = 10, b = \frac{3}{2}\pi$.

(c) $\sigma = 2, \alpha = 1, b = 0.4$.

(d) $\sigma = 5, \alpha = 10, b = -0.8$.

Figure 2.5: (a)–(b): Gradient orientation histogram bin values for bin centered at $b = \frac{3}{2}\pi$ with a tonal range $\beta = 0.62$. (c)–(d): Shape index histogram values for bins centered at $b = 0.4$ and $b = -0.8$ with tonal range $\beta = 0.26$.

Rather than the image descriptors above, a slightly different way of representing an images can be achieved using image patches. While the basic concept of localized information is similar, image patches are often more useful in the

context of reconstructing an image. This is different than above, where we primarily seek a descriptive representation and not so much a generative. Image patches as local descriptors will therefore be a subject of some separate attention now.

2.2.3 Patch based methods

Image patches are small, usually quadratic, excerpts of an image. As such an image patch captures local texture and intensity distribution. The distribution of patches extracted from an image can therefore be used as a representation of said image. Furthermore, patches extracted from multiple images collectively represent these images. However, a large number of patches might be necessary to represent all the characteristics found in a collection of images; how do we extract 'the useful' ones? Or put in other terms: How to select a subset of features (patches) that best represent, reconstruct or discriminate in a specific instance? Enter compressed sensing/sparse coding/dictionary learning (Donoho, 2006, Elad, 2010, Mairal et al., 2008b).

In dictionary learning the goal is to represent a set of signals $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$, $\mathbf{x}_i \in \mathbb{R}^p$ by an overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{p \times k}, k > p$ and a sparse weighting for each observation $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]^T, \mathbf{a}_i \in \mathbb{R}^k$. In a patch-based context the signals $\mathbf{x}_i$ are vectorized patches, illustrated in Figure 2.6. Sparsity is usually induced on the $\mathbf{a}_i$'s by an ℓ_1-norm penalty, since the "ℓ_0-norm" is not convex (and thus not a real norm) and thus requires more involved algorithms (Mairal et al., 2007, 2008a). The introduction of the Lasso by Tibshirani (1996) and least angle regression (LARS) by Efron et al. (2004) brought fast algorithms for solving ℓ_1-norm regularized problems for fixed $\mathbf{D}$. Mairal et al. (2008b) solves Eq. (2.30) jointly for $\mathbf{D}$ and $\mathbf{A}$,

$$\arg\min_{\mathbf{A},\mathbf{D}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_1, \tag{2.30}$$

where λ is a sparsity inducing regularization parameter due to it acting on the ℓ_1-norm. Mairal et al. (2010) furthers this with an online learning algorithm that scales well with large data sets.

Mairal et al. (2008a) use dictionaries for supervised classification and incorporates the dictionary's discriminative ability into the optimization, i.e., ensures that a dictionary good for representing one class is simultaneously bad at other classes. Dahl and Larsen (2011) approaches this problem by modelling a dictionary in two linked spaces; one in the observed image space and one in the label

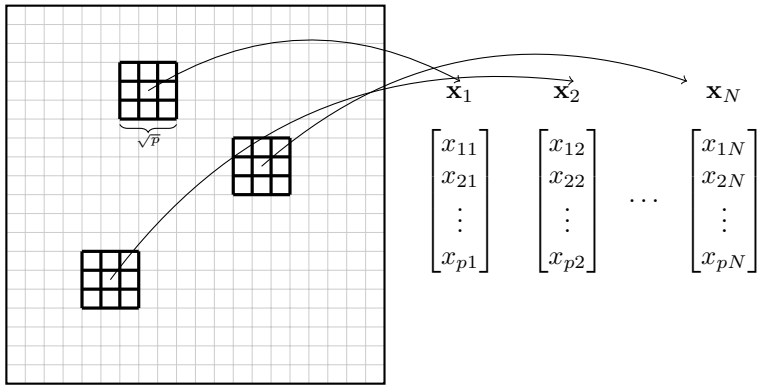


Figure 2.6: Extraction of patches of size $\sqrt{p} \times \sqrt{p}$ from an image. Each patch is stacked to a feature vector of length p. Note the similarity to Figure 2.3.

space. Optimization of the dictionary proceeds to move dictionary atoms further away based on the distances in *label* space rather than *intensity* space. It should be noted that Dahl's approach is based on each patch being represented by a single atom (cluster), i.e., $\|\mathbf{a}_i\|_0 = 1$ which can be seen as the ultimate sparsity. This reduces the implementation to a modified iterated 1-nearest neighbor algorithm. Mairal et al. (2012) presents a general formulation of supervised dictionary learning, algorithms for solving the yielded optimization problem, and how to adapt the formulation to a wide variety of tasks.

The reconstructive ability of learned dictionaries have also proven useful for constructing super resolution images (Yang and Wright, 2010, Wang et al., 2012). Image patches are also the basic component for observations in many computer vision tasks (see e.g., Cristinacce and Cootes, 2008, Cherian et al., 2014) and in deep learning of image structure (Hinton et al., 2006, Jarrett et al., 2009, Salakhutdinov and Hinton, 2012). Patches are particular popular in the latter case, as deep learning of image features requires large amounts of training data, which is not always available; however, a multitude of patches in various orientations, scales, etc. can be extracted from a single image thus augmenting the data set.

We will now turn to probabilistic modelling with graphical models. This is useful for the above mentioned disciplines, where a patch is observed with some probability or the probability distribution of a set of features need to be learned. Further, these models can be used to jointly model observed data in conjunction with the prior expectation of the (geometric) structure.

2.3 Modelling geometry

Markov random fields (MRFs) were introduced in the 1980s (e.g., Hassner and Sklansky, 1980, Geman and Geman, 1984, Ripley, 1991) as a class of stochastic processes to model both the prior and posterior distribution of an image. The work related to this thesis is mostly concerned with Gaussian MRFs for fitting geometric models to images. First, MRFs in general will be introduced here with the primary sources being Geman and Geman (1984), Carstensen (1992), Bishop (2007), Li and Kanade (2009).

Consider a graph with nodes (or sites) $\mathbf{s} = \{s_1, \ldots, s_n\}$. A neighborhood system $\mathcal{N} = \{N_i, s_i \in S\}$ is a collection of subsets of the n sites for which $s_i \notin N_i$ and $s_j \in N_i \Leftrightarrow s_i \in N_j$, where N_i are the neighbors of s_i. As such the neighborhood system defines the edges of the graph. A clique is a subset of sites in which every pair of sites are neighbors and a maximal clique is a subset of nodes that are fully connected. The notation $i \sim j$ means that sites s_i and s_j are neighbors. Examples of such a configurations can be seen in Figure 2.7.

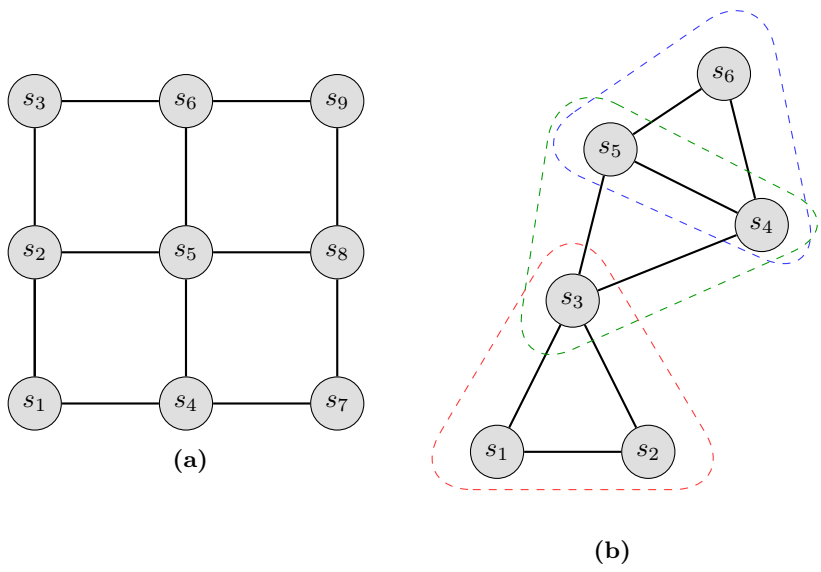


Figure 2.7: (a) A regular MRF, e.g., a pixel grid with the neighborhood defined as the horizontal and vertical neighbors. (b) An irregular neighborhood structure. One-cliques are the same as the nodes, two-cliques are every pair of nodes connected and three-cliques (the maximal cliques for this graph) are marked in dashed lines.

Each site in an MRF is associated with a random variable X_s with values $x_s \in \Omega$. The sample space Ω depends on the problem, e.g., in a binary labelling problem $x \in \{0, 1\}$ and Ω is the set of all 2^n possible configurations. The random variable associated with site s can also be referred to as X_s with value x_s.

Definition 2.3 (Markov random field). A random field X is a Markov random field with respect to $\mathcal{N} = \{N_i, s_i \in S\}$ if and only if

1. $P(\mathbf{X} = \mathbf{x}) > 0 \; \forall \; \mathbf{x} \in \Omega$
2. $P(X_i = x_i | X_j = x_j, i \neq j) = P(X_i = x_i | X_j = x_j, s_j \in N_i) \; \forall \; i \in \{1, \ldots, N\}$.

▲

2.3.1 MRFs are undirected graphical models

Probabilistic graphical models let the nodes in a graph represent variables and the edges specify dependence properties such that the entire graph specify the joint distribution $p(\mathbf{x}) = p(x_1, \ldots, x_n), x \in \Omega$. There are two main classes of probabilistic graphical models: directed and undirected.

Directed graphical models specify a factorization of the joint distribution $p(\mathbf{x})$ over the set of variables $\mathbf{x} = \{s_1, \ldots, s_p\}$ into a product of local conditional distributions

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | \mathrm{pa}_i) \tag{2.31}$$

where pa_i is the set of parents of the i'th node in a graphical model representing the joint distribution (Bishop, 2007), see Figure 2.8 for an example. In contrast, MRFs are undirected probabilistic graphical models which factorize the joint distribution into a product over the maximal cliques of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}_m} V_C(\mathbf{x}_C) \tag{2.32}$$

where $\mathcal{C}_m$ is the set of maximal cliques in the graph, $\mathbf{x}_C$ the set of variables in a clique C and $Z = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}_m} V_C(\mathbf{x}_C)$ is the partition function, ensuring that the probability over the variables sums to one. We can write this factorization in terms of the maximal cliques without loss of generality, since cliques must be subsets of the maximal cliques. The function V_C depends only on the neighbors $\mathcal{N}_C$ and is called a *potential* or *potential function*.

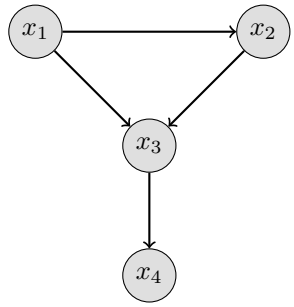


Figure 2.8: A directed graph representing the joint distribution $p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3)$.

Comparing with directed graphical models, there is not the same equivalence between the local dependence properties and conditional independence. However, by considering only non-negative potential functions (and thus ensuring $p(\mathbf{x}) \geq 0$) the Hammersley-Clifford theorem (Hammersley and Clifford, 1971) establishes an equivalence between distributions factorizable into cliques as in Eq. (2.32) with distributions with conditional independences that can be established from a graph (Bishop, 2007), i.e., a connection between the local property (Markovianity) of MRFs with the global property of Gibbs random fields (GRFs) (Li and Kanade, 2009). This relation allows us to specify MRFs in terms of potentials, which is much easier than trying to specify local characteristics for a global configuration (Li and Kanade, 2009). The non-negativity $V_C(s) \geq 0$ makes it convenient to formulate potential functions as Boltzmann distributions

$$V_C(\mathbf{x}_C) = \exp(-U(\mathbf{x}_C)) \tag{2.33}$$

where $U : \Omega \mapsto \mathbb{R}$ is the *energy function*. The joint distribution is obtained from Eq. (2.32) as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \exp(-U(\mathbf{x}_C)) = \frac{1}{Z} \exp\left(- \sum_{C \in \mathcal{C}} U(\mathbf{x}_C) \right) \tag{2.34}$$

where the factorization over cliques conveniently reduces to a sum over energies. With the inclusion of a temperature T

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{1}{T} \sum_{C \in \mathcal{C}} U(\mathbf{x}_C) \right) \tag{2.35}$$

this is equivalent to the Gibbs distribution introduced to express the probability of a (physical) system being in a state with a certain energy (Carstensen, 1992).

2.3.2 Ising, Potts and Gaussian auto-models

The simplest MRF model is the binary (two-state) Ising model (Ising, 1925). Introduced in 1925 in statistical physics by Ising, treated in statistics in the 1960s and in image analysis in the 1980s it has become the binary segmentation model of choice with the publicly available *graph-cut* algorithm by Boykov and Kolmogorov (2004). The graph-cut algorithm exploits the min-cut/max-flow equivalence from graph theory and ensures fast, global convergence to a minimum energy state. See Section 2.3.3 for more about energy minimization.

The general form of the energy function is

$$U(x) = \sum_{s \in S} u_1(x_s) + \sum_{s \in S} \sum_{r \in N_s} u_2(x_s, x_r) \tag{2.36}$$

where $u_1(\cdot)$ and $u_2(\cdot, \cdot)$ are energy functions over one- and two-cliques respectively. The Ising model is the special case in which x is binary, i.e., $\Omega = \{0, 1\}$ and

$$U(x) = \alpha \sum_{s \in S} x_s + \beta \sum_{s \sim r} x_s x_r \ . \tag{2.37}$$

for parameters α and β, where β controls the neighborhood (or bonding) strength.

The Potts (Potts, 1952, Wu, 1982) model extends the Ising model to more than two states and is usually optimized using the *alpha-expansion* heuristic (Boykov et al., 2001), which cannot ensure a global optimum. An exception of this is for ordered labels, in which case the multistate problem can be solved exact (Li and Kanade, 2009). The energy function can be written as the energy of site s having state k conditional on the value of the neighbors of s as

$$U(x_s = k | N_s) = \alpha_k + \beta_k u_s(k) \tag{2.38}$$

where $u_s(k)$ is the number of neighbors of s having value k. Other variations of the Potts model exists, see e.g. Carstensen (1992) for examples.

The above models are also referred to as *auto-models*, since they are formulated in terms of pair-wise interactions. When the number of possible states are the real values and the joint distribution is a multivariate normal an auto-model is called a Gaussian MRF (GMRF). The energy for a site s conditional on its neighbors is

$$U(x_s | N_s) = -\frac{1}{2\sigma^2} \left(x_s - \mu_s - \sum_{r \sim s} \beta_{s,r} (x_r - \mu_r) \right)^2 \tag{2.39}$$

with known partition function $Z = \frac{1}{\sqrt{2\pi\sigma^2}}$.

2.3.3 MCMC sampling for energy minimization

Determining the parameters for minimum energy in an MRF, i.e. maximize the posterior distribution of the parameters $\mathcal{G}$, is non-trivial in most cases due to the state space often being prohibitively large. E.g., for an image of size 128×128 with three possible labels the number of possible configurations (the size of Ω) is $128^6 \approx 4.4 \cdot 10^{12}$.

Further, GMRFs cannot be solved to a global optimum using, e.g., the graph-cut algorithm, but need to be solved using a sampling approach. GMRFs are of particular interest in Section 2.3.4 where the spatial position of a node is assumed to be at an approximate distance from its neighbors. Therefore Markov Chain Monte Carlov (MCMC) sampling methods will be introduced here following that of Bishop (2007).

The method of iterated conditional modes (ICM) (Kittler and Föglein, 1984) is the simplest way to minimize the energy in an MRF. It can be seen as a element-wise steepest descend algorithm, i.e., each site s is visited in order keeping all other sites fixed at their value. The value x_s is then changed in direction of maximum energy loss, i.e., the steepest descend direction. This is guaranteed to minimize the global energy or leave it unchanged. This is repeated until convergence. The danger of ICM is obviously that it is likely to converge to a local rather than the global minimum.

Markov Chain Monte Carlo (MCMC) is a family of methods that scales better with high dimensions. Similar to other sampling methods a proposal distribution $q(\mathbf{x})$ is used for assessment of newly drawn samples. However, for MCMC methods the proposal distribution is dependent on the current state $\mathbf{x}^\star$ and the sequence of samples $\mathbf{x}_1, \mathbf{x}_2, \ldots$ forms a Markov Chain.

2.3.3.1 Markov chains

A first order Markov chain is defined such that the conditional independence holds

$$q(\mathbf{x}_l|\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}) = q(\mathbf{x}_l|x_{l-1}) \ , \tag{2.40}$$

i.e., such that a state is only dependent on the previous state. The Markov chain can be specified by the pdf of an initial state $q(\mathbf{x}_0)$ together with transition probabilities specifying the conditional probabilities for subsequent samples

$$T_l(\mathbf{x}_l, \mathbf{x}_{l+1}) \equiv q(\mathbf{x}_{l+1}|\mathbf{x}_l) \ . \tag{2.41}$$

If the transition probabilities are the same for all l the Markov chain is called homogeneous. *Detailed balance* of the transition probabilities is defined as

$$q(\mathbf{x})T(\mathbf{x}', \mathbf{x}) = q(\mathbf{x}')T(\mathbf{x}', \mathbf{x}) \tag{2.42}$$

and ensures invariance of q. A Markov chain with transition probabilities that respect detailed balance is said to be reversible.

A Markov chain set up such that the sampled distribution is invariant ensures that we sample from this distribution. However, the property of *ergodicity* must also hold, meaning that as $l \to \infty$ the distribution $q(\mathbf{x}_l) \to p(\mathbf{x})$ irrespective of the choice of inital state $\mathbf{x}_0$ (Bishop, 2007).

2.3.3.2 Metropolis-Hastings

The Metropolis-Hastings algorithm is an MCMC sampling method where a sample $\mathbf{x}^\star$ is accepted with probability

$$p_{\text{accept}}(\mathbf{x}^\star, \mathbf{x}_l) = \min\left(1, \frac{\tilde{p}(\mathbf{x}^\star)q(\mathbf{x}_l|\mathbf{x}^\star)}{\tilde{p}(\mathbf{x}_l)q(\mathbf{x}^\star|\mathbf{x}_l)}\right) . \tag{2.43}$$

when the current state is $\mathbf{x}_l$ and we consider probability density functions of the form $p(\mathbf{x}) = \frac{1}{Z}\tilde{p}(\mathbf{x})$, i.e., we explicitly do not consider the partition function. This can be shown to satisfy detailed balance by

$$\begin{aligned} p(\mathbf{x})q(\mathbf{x}|\mathbf{x}')p_{\text{accept}}(\mathbf{x}', \mathbf{x}) &= \min\left(p(\mathbf{x})q(\mathbf{x}|\mathbf{x}'), p(\mathbf{x}')q(\mathbf{x}'|\mathbf{x})\right) \\ &= p(\mathbf{x}')q(\mathbf{x}'|\mathbf{x})p_{\text{accept}}(\mathbf{x}, \mathbf{x}') \end{aligned}$$

Thus the Markov chain specified by the Metropolis-Hastings transition probability specifies an invariant distribution $p(\mathbf{x})$. For $q(\mathbf{x}'|\mathbf{x}) = q(\mathbf{x}|\mathbf{x}')$ this reduces to the standard Metropolis sampling method, which therefore also satisfies detailed balance (Bishop, 2007).

2.3.3.3 Gibbs sampling

Gibbs sampling was introduced in the seminal paper by Geman and Geman (1984) and can be seen as a special case of the Metropolis-Hastings algorithm (Bishop, 2007). This sampling method exploits that it can be easier to sample the value for a single site conditional on all the rest $p(x_i^l|z_j^{l-1}, j \neq i)$, rather than sample $p(\mathbf{x}^l)$ directly. Theoretically, when sampling for infinity the Gibbs sampler will eventually sample from the true distribution.

A problem with Gibbs sampling is the autocorrelation of subsequent samples. Simulated annealing also introduced by Geman and Geman (1984) is one approach to alleviate this. In a simulated annealing scheme, the temperature T in Eq. (2.35) is lowered from some initial temperature T_0 towards 0 according to a temperature scheme, e.g., $T(l) = cT(l-1)$. For high temperatures, changes that increase the energy are more likely to get accepted than for low temperatures. I.e., larger, more improbable, steps are allowed in the beginning of such a scheme which can avoid converging to local minima. When the temperature is low, only energy-lowering moves are accepted.

2.3.3.4 Coding schemes for practical sampling

The naive approach to suggesting new states in an MCMC sampling scenarion would be to visit every site one-by-one in random order, while keeping values at all other sites fixed. In practice a coding scheme is usually employed, where the sites are divided into *colors* such that no neighboring sites can be of the same color. Due to the Markovian property sites of the same color can then be updated simultaneously, rather than choosing a single site at a time.

An optimal coloring uses the minimum possible number of colors called the *chromatic number*. In many cases, the neighborhood structure is simple and the coding scheme is known in advance. However, for less simple neighborhood structures, the problem of determining the coding scheme is known as the graph coloring problem. While determining the chromatic number is an NP-complete problem and thus infeasible to determine, the Welsh-Powell algorithm (Welsh and Powell, 1967, Kubale, 2004) can be useful for obtaining a sub-optimal solution: 1) Order the vertices in the graph according to degree, 2) visit each vertex sequentially according to this ordering and 3) assign the smallest color number not in use by any of the vertex neighbors. This simple, but practical, algorithm has been used for the work in Papers E and F.

2.3.4 Decoupling observation and geometry

The models considered above integrate some prior knowledge $p(\mathbf{s})$ about the structure modelled and relates it to the observed image through an observation model $p(\mathbf{x}|\mathbf{s})$ by Bayes rule, usually written in terms of the unnormalized joint posterior

$$p(\mathbf{s}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{s})p(\mathbf{s}) \; . \tag{2.44}$$

The most common models define the prior $p(\mathbf{s})$ on the same lattice as the observed image (Carstensen, 1996), such that there is one-to-one correspondence

between sites and pixels. Geman and Geman (1984) introduced the line process as a way to decouple the prior from the image lattice and thus potentially making it smoother. The models considered here are primarily motivated by Carstensen (1996) and Hartelius and Carstensen (2003) and are similar to deformable template models (Amit and Kong, 1996, Jain and Lakshmanan, 1996, Güdükbay et al., 1997, Jain et al., 1998), which are strongly related to active shape and appearance models (Cootes et al., 1995, 2001, Cristinacce and Cootes, 2008).

Here we will consider the problem of fitting a regular lattice to an image, where the lattice geometry is specified by the prior, separately from the pixel grid, and the likelihood of the lattice will be specified through an observation model. The problem of fitting a grid structure to an image is concerned with determining the node positions $\mathbf{s}_i = \{x_i, y_i\}, i = 1, \ldots, N$ of N nodes. A graphical model of a regular triangular lattice in such a scenario is sketched in Figure 2.9. Note the directed arrows towards the observed image, indicating the conditional dependence of $p(\mathbf{x}|\mathbf{s})$ on $\mathbf{s}$, and the undirected connections in the grid structure. The undirected graphical model for the geometric prior is a good way of specifying grid priors, since grids are often repetitive patterns of local models. In these cases it also means that there is a Markov property in the grid prior as introduced previously.

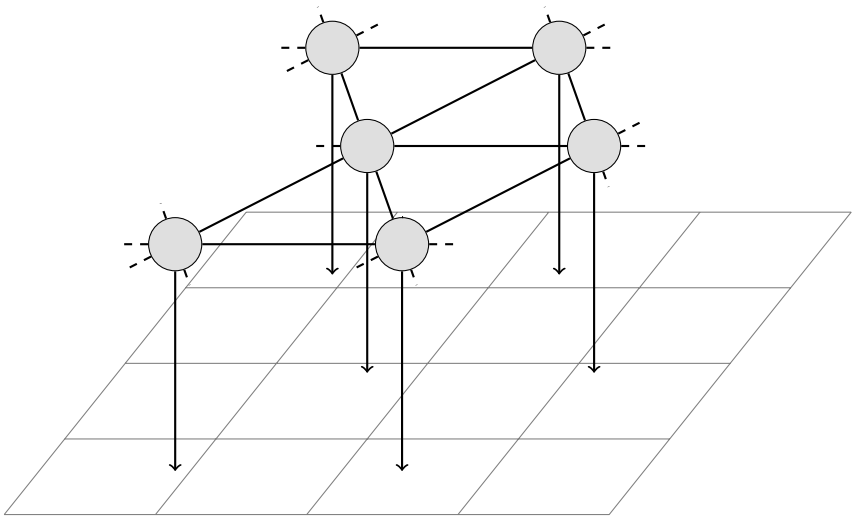


Figure 2.9: Decoupling of the geometric structure from the image lattice is illustrated with a triangular lattice. The probability of the structure $\mathbf{s}$ given the image $\mathbf{x}$ can be written as $p(\mathbf{s}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{s})p(\mathbf{s})$ according to this graphical model.

Consider a regular grid with a template node distance t. The distance d_{ij} from a node s_i to its neighbors $s_j, i \sim j$ is then expected to be

$$d_{ij} = t + \varepsilon \quad , \; \varepsilon \sim \mathcal{N}(0, \sigma_t^2) \tag{2.45}$$

where σ_t is a parameter controlling the slack of the position. This is illustrated in Figure 2.10 for various grids.

Consider now a template node distance t_{ij} defined locally, i.e., the distance d_{ij} is expected to be the approximately the same as the lengths of the neighboring N_{ij} edges $d_{k\ell}, k\ell \sim ij$:

$$d_{ij} = t_{ij} + \varepsilon \tag{2.46}$$

where

$$t_{ij} = \frac{1}{N_{ij}} \sum_{ij \sim k\ell} d_{k\ell} \; . \tag{2.47}$$

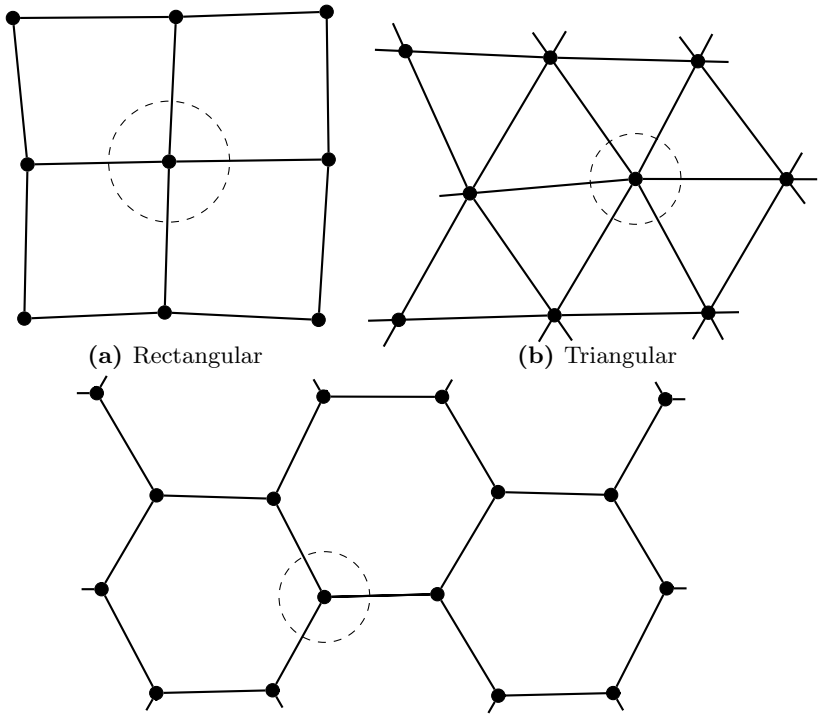
This requires definition of an edge neighborhood, e.g., all edges connected to the same vertex are considered neighbors or a directionality dependent scheme as suggested by Hartelius and Carstensen (2003).

Having a model for the distance allows us to set up an energy function for a node position, given the position of its neighbors as

$$U_{\text{grid}}(\mathbf{s}_i | N_i) = \frac{1}{2\sigma_t^2} \sum_{i \sim j} \left(\|\mathbf{s}_i - \mathbf{s}_j\|_2 - d_{ij} \right)^2 \; , \tag{2.48}$$

which is seen to be a GMRF. The energy in such a GMRF can be minimized by alternating between determining d_{ij} and taking steps in $\mathbf{s}$ using, e.g., Metropolis-Hastings. This is the basic model used for fine adjusting grid structures for the papers in Chapter 6.

The observation model is very flexible, in that it simply needs to map the given position of a node to an observation energy. The energy field could be the result of an image filtering operation, e.g., edge enhancement or blob detection, a probability map from a segmentation procedure or anything else where low energy is proportional to the likelihood of the position. More general energy functions, edge priors and observation models can be found in Hartelius and Carstensen (2003).



(a) Rectangular **(b)** Triangular

(c) Hexagonal

Figure 2.10: Various regular lattices, where all edges in a noise free lattice have the same length. Note that the triangular and hexagonal lattice are each others dual lattices. The dashed lines indicate the modelling of uncertainty of a node position given its neighboring nodes.

Chapter 3

Manifold learning

3.1 Linear decomposition

A common task when dealing with multivariate images is decomposition of the image into its "interesting" signals. This could be for visual inspection or as a pre-processing step for further analysis; it could be for a single image (one set of variables), two images (two-set) or multiple images (multi-set). The definition of "interesting" depends on the decomposition method. A brief summary of some standard one-set image decomposition methods will be given here. We will mainly be concerned with methods not using a fixed set of basis functions, in contrast to, e.g., Fourier transforms or wavelets (Gershenfeld, 1999).

In linear decomposition the model for an observed set of signals $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ is

$$\mathbf{X} = \mathbf{Z}\mathbf{A} + \varepsilon, \tag{3.1}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_q]$ are the true underlying signals, $\mathbf{A} \in \mathbb{R}^{q \times p}$ is a projection matrix and ε is noise. The projection matrix is also referred to as the *mixing* or *basis change* matrix. The goal is usually to simultaneously determine $\mathbf{A}$ and $\mathbf{Z}$ subject to some measure of optimality.

Linear single-set image decomposition is in very general terms defined in Definition 3.1.

Definition 3.1 (Image decomposition). Image decomposition seeks the linear combination $\mathbf{a}_i \in \mathbb{R}^p$ of the image data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, where $N = mn$, as the solution to

$$\arg\max_{\mathbf{a}_i} g(\mathbf{X}\mathbf{a}_i)$$

subject to some condition on the independence between $\mathbf{a}_i$ and $\mathbf{a}_j, j = 1, \ldots, i-1$ for $i > 1$. The q found components are sorted such that $g(\mathbf{X}\mathbf{a}_i) > g(\mathbf{X}\mathbf{a}_{i-1})$ for $i > 1$. Here $g : \mathbb{R}^N \mapsto \mathbb{R}$ is an arbitrary function, which will vary between different image decomposition methods. Note that this definition is only useful in the simple cases of non-regularized and non-constrained methods. ▲

This definition can encompass some of the classical image decomposition methods: principal components analysis (PCA) can be formulated with $g(\cdot) \equiv \text{var}(\cdot)$ and maximum autocorrelation factor (MAF) analysis uses

$$g(\mathbf{X}(\mathbf{x})\mathbf{a}_i) \equiv \text{corr}\left(\mathbf{X}(\mathbf{x})\mathbf{a}_i, \mathbf{X}(\mathbf{x}+\Delta)\mathbf{a}_i\right)$$

where Δ is a spatial displacement (Switzer and Green, 1984). MAF analysis is a special case of the minimum noise fraction (MNF) analysis which maximizes the signal-to-noise ratio $g(\mathbf{X}\mathbf{a}_i) \equiv \frac{\mathbf{a}_i \mathbf{S}_S \mathbf{a}_i}{\mathbf{a}_i \mathbf{S}_N \mathbf{a}_i}$ using a noise model to estimate the noise and signal covariance matrices $\mathbf{S}_N$ and $\mathbf{S}_S$ (Green et al., 1988, Nielsen, 2011). These are all examples of variance preserving decompositions, where the condition mentioned in Definition 3.1 is orthogonality. Further they all assume approximately Gaussian distributed data (or noise).

We distinguish between orthogonal and oblique decomposition methods. The first type of methods yields an orthogonal $\mathbf{A}$, and thus $q \leq p$ and for $q = p$ the matrix $\mathbf{A}$ specifies a rotation of the original space. Oblique decomposition methods can yield an arbitrary number of projection directions. Embedding a new observation $\mathbf{x}_{\text{new}}$ into a subspace is simply

$$\mathbf{z}_{\text{new}} = \mathbf{A}\mathbf{x}_{\text{new}} \,. \tag{3.2}$$

However, as noted by Shen and Huang (2008) care should be taken when calculating, e.g., the variance contained in a q-dimensional non-orthogonal subspace. For an orthogonal subspace of dimension p the variance can be calculated as a sum of the variance in each projection direction as $\text{var}(\mathbf{Z}\mathbf{A}) = \sum_{i=1}^{p} \text{var}(\mathbf{Z}\mathbf{a}_i^T)$. When the projections are not orthogonal, there may be overlap in information between the projection directions and the variance for the data in the q-dimensional subspace defined by $\mathbf{A}_q$ needs to be calculated using the principle of an oblique projection as

$$\text{var}(\mathbf{Z}\mathbf{A}_q) = \text{tr}(\mathbf{X}_q^T \mathbf{X}_q)$$

where

$$\mathbf{X}_q = \mathbf{X}\mathbf{A}_q(\mathbf{A}_q^T\mathbf{A}_q)^{-1}\mathbf{A}_q^T \,.$$

This is important for, e.g., sparse principal component analysis and was not noted in the original paper by Zou et al. (2006).

For many image decomposition methods it is common to assume that the variables are normally distributed; this is in many cases a fair assumption, e.g., when dealing with optical images as mentioned in Section 2.1. Assuming Gaussianity makes it sufficient to consider second-order moments of the distribution thus simplifying optimization.

3.1.1 Information theoretical

In Section 2.1.3 entropy was introduced as an information theoretical measure useful for quantifying the amount of "surprise" or information content in a distribution. Entropy has its origin in thermodynamics, and we saw in Section 2.3.1 how the Gibbs distribution is important in the context of MRFs. Naturally, methods for optimizing measures from statistical thermodynamics and information theory exist.

Independent component analysis (ICA) constitutes a class of methods searching for latent non-Gaussian variables in the data. Non-Gaussianity is claimed to be a better measure for "signal" in many real-world scenarios (Hyvärinen et al., 2001). The most common variant of ICA is Infomax by Bell and Sejnowski (1995), where non-Gaussianity is enforced by assuming a particular heavy-tailed distribution of the latent variables under the model $\mathbf{z} = g(\mathbf{u}), \mathbf{u} = \mathbf{A}\mathbf{x} + \mathbf{a}_0$, where $\mathbf{A}$ is the mixing matrix and $\mathbf{x}$ the observed data. This is usually accomplished by introducing a sigmoidal transfer function $g(u) = \frac{1}{1+\exp(-u)}$. Since the introduction of ICA, a variety of models and solutions have emerged (see e.g. Hyvärinen et al., 2001, Shwartz et al., 2005), some of which are formulated in terms of probabilistic graphical models and thus probabilistic in nature (e.g., Hinton et al., 2001, Chan et al., 2003, Bishop, 2007).

The information theoretical measure *mutual information* was introduced in Section 2.1.3.2 and is the foundation for the two-set decomposition method developed in Paper A. Two-set decomposition will be introduced here and the contribution of Paper A will be summarized in Chapter 4.

3.1.2 Two-set decomposition

Two-set decomposition can be defined as in Definition 3.2.

Definition 3.2 (Two-set image decomposition). For the two sets of variables $\mathbf{X} \in \mathbb{R}^{N \times p}$ and $\mathbf{Y} \in \mathbb{R}^{N \times q}$, $q \leq p$ two-set decomposition seeks pairs of linear combinations $(\mathbf{a}_i, \mathbf{b}_i)$ as the solution to

$$\arg\max_{\mathbf{a}_i, \mathbf{b}_i} g(\mathbf{X}\mathbf{a}_i, \mathbf{Y}\mathbf{b}_i)$$

subject to some condition on the independence between $\mathbf{a}_i$ and $\mathbf{a}_j, j = 1, \ldots, i-1$ for $i > 1$ and between $\mathbf{b}_i$ and $\mathbf{b}_j, j = 1, \ldots, i-1$ for $i > 1$. Here $g : (\mathbb{R}^N, \mathbb{R}^N) \mapsto \mathbb{R}$ defines a measure of association, which will vary between decomposition methods. ▲

Canonical correlation analysis (CCA) by Hotelling (1936) is the best-known two-set decomposition method and is described in various textbooks (see e.g., Anderson, 1984, Wackernagel, 1995).

As suggested by the name CCA maximizes the correlation ρ between linear combinations of these two sets of variables ($\mathbf{X}$ and $\mathbf{Y}$):

$$\rho = \operatorname{corr}\left\{\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right\} = \frac{\mathbf{a}^T \boldsymbol{\Sigma}_{12} \mathbf{b}}{\sqrt{\mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a}}\sqrt{\mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b}}} \tag{3.3}$$

where $\boldsymbol{\Sigma}_{12} = \operatorname{cov}(X, Y)$. This can be done by solving the generalized eigenvalue problem

$$\rho^2 = \frac{\mathbf{a}^T \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \mathbf{a}}{\mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a}} = \frac{\mathbf{b}^T \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \mathbf{b}}{\mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b}} \tag{3.4}$$

where the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_q$ with corresponding eigenvalues $\rho_1^2 \geq \ldots \geq \rho_q^2$ are the desired projection directions for $\mathbf{X}$. For more details, see Hotelling (1936), Nielsen (2002). Two-set decomposition is of interest in relation to Paper A.

3.2 Locality-based embedding

Locality-based manifold learning is motivated by a desire to have the manifold tend towards a certain structure, e.g., that near-by points in input space are also near-by in the learned feature space. Here we will consider methods using a weighted graph, where the presence/abscence of an edge and its weight regularizes the solution towards the desired structure. Isomap (Tenenbaum et al., 2000), locally linear embedding (LLE) (Roweis and Saul, 2000), locality preserving projections (LPP) (He and Niyogi, 2003) and Laplacian eigenmaps (Belkin and Niyogi, 2003) are different algorithms proposed to solve this problem. Yan

et al. (2007) proposed a unified framework based on a general graph formulation, where all of these methods can be seen as special cases with different criteria for setting up the graph. This graph embedding framework not only encapsulates these methods (and others, such as PCA and LDA), but it does so linearly and thus does not suffer from the same weakness as Isomap, LLE and Laplacian eigenmaps, namely that they are only defined in terms of the training points and embedding of new test is not well defined (Yan et al., 2007, Cai et al., 2007b).

Define a graph G with N vertices, one for each data point $x_i, i = 1, \ldots, N$, a symmetric matrix $\mathbf{W}$ with W_{ij} the weight of the edge joining x_i and x_j and the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with the degree of each vertex, i.e., $D_{ii} = \sum_j W_{ij}$. The "graph-preserving criterion" for the linear embedding $\mathbf{x} \mapsto \mathbf{z}$ is then

$$\underset{\substack{\mathbf{a} \\ \text{s.t. } \mathbf{a}^T\mathbf{X}^T\mathbf{B}\mathbf{X}\mathbf{a}=d}}{\arg\min} \sum_{i \neq j} \|\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j\|^2 W_{ij} = \arg\min \frac{\mathbf{a}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{a}}{\mathbf{a}^T\mathbf{X}^T\mathbf{B}\mathbf{X}\mathbf{a}} \tag{3.5}$$

where $d \in \mathbb{R}$ and $\mathbf{B} \in \mathbb{R}^{N \times N}$ is a constraint matrix such that $\mathbf{a}^T\mathbf{X}^T\mathbf{B}\mathbf{X}\mathbf{a} = d$ either serves the purpose of avoding trivial solutions, normalizes the solution or guides the solution by defining separate graph weights penalizing proximity, instead of encouraging proximity as in $\mathbf{W}$. Yan et al. (2007) provides graph versions of the aforementioned methods, as well as linearization and kernelization techniques for generalizing the found solution to new samples.

The notion of locality preserving embeddings proves useful for semi-supervised methods. Semi-supervised methods are concerned with situations where a part of the data set (usually the minority) is labelled, while a – potentially significant – amount of data is unlabeled. While the unlabeled data are not useful for the supervised part of the task, e.g., discriminant analysis, the resulting manifold should still learn something from them. This could for instance be that the manifold honors a locality criterion as above. Cai et al. (2007a) and Song et al. (2008) simultaneously proposed semi-supervised discriminant analysis in this context with the optimization problem posed as

$$\arg\max_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{S}_B\mathbf{a}}{\mathbf{a}\mathbf{S}_T\mathbf{a} + \alpha J(\mathbf{a})} \tag{3.6}$$

where $\mathbf{S}_B$ and $\mathbf{S}_T$ are the between classes scatter and the total covariance, α is a regularization parameter and $J(\mathbf{a}) = \sum_{ij}(\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j)S_{ij}$ with $\mathbf{S}$ being the edge matrix linking nearby observations, e.g., with a binary link to the k

nearest neighbors. Defining the block diagonal matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & & \\ & \ddots & & \\ & & \mathbf{W}_c & \\ & & & \mathbf{0} \end{bmatrix}$$

where $\mathbf{W}_j$ is a matrix with all elements equal to $1/N_j$ where N_j is the number of observations in the j'th class. Further we define

$$\tilde{\mathbf{I}} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \tag{3.7}$$

where $\mathbf{I}$ is an $N_\ell \times N_\ell$ identity matrix and $N_\ell = \sum_{j=1}^{c} N_j$. The optimization problem from Eq. (3.6) can then rephrased as

$$\arg\max_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{a}}{\mathbf{a}^T\mathbf{X}^T(\tilde{\mathbf{I}} + \alpha\mathbf{L})\mathbf{X}\mathbf{a}} , \tag{3.8}$$

where the observations must be sorted according to class id, with the unlabeled observations in the end, to align with $\mathbf{W}$ above. $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the graph Laplacian as before with $\mathbf{S}$ defining the local structure. Note how the between classes scatter is only influenced by labelled observations, while the penalty term in the denominator incorporates all observations. This is seen to fit into the framework from Eq. (3.5).

Interesting work is also being carried out for learning embeddings using neural networks, based solely on neighborhood relationships (see e.g., Hadsell et al., 2006).

3.3 Non-linearity via kernel methods

Kernel methods in machine learning and pattern analysis were introduced as a canonical framework for modelling unknown non-linear relations in data. A naive approach to expanding a basis to include non-linear relations is to expand a set of variables $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ with for instance quadratic relations such that $\tilde{\mathbf{X}} = \{\mathbf{x}_1, \mathbf{x}_1^2, \ldots, \mathbf{x}_p, \mathbf{x}_p^2\}$. In case any of the variables are better represented squared this could be a better basis for pattern analysis. However, how do we know whether to include quadratic relations and not cubic, square roots or something even more exotic? Kernel methods can be seen as a framework to introduce more general non-linearities than a manual basis expansion can

provide. This introduction to kernel methods will follow that of Abrahamsen (2009) and Shawe-Taylor and Cristianini (2004).

Kernel methods consist of two stages: First, a mapping from the input space $\mathcal{X}$ to a potentially non-linear feature space $\mathcal{H}$. Secondly, some linear pattern analysis method is applied in this feature space. This is sketched for a discriminative scenario in Figure 3.1.

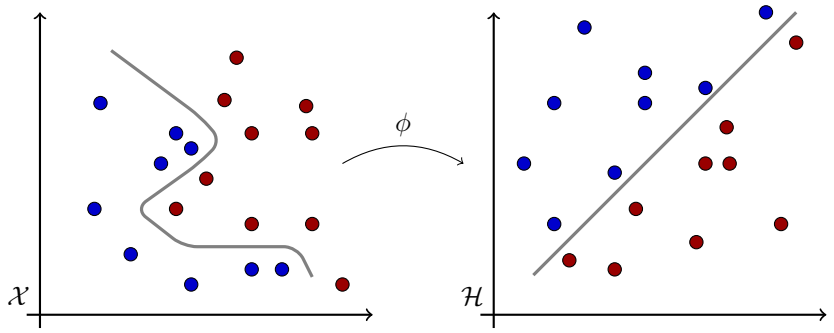


Figure 3.1: Idealized sketch of the principle of kernel embedding for a discriminative task. While a non-linear decision boundary is optimal in the input space, for some non-linear embedding ϕ the optimal boundary is linear. Similar to sketches by, e.g., Abrahamsen (2009) and Shawe-Taylor and Cristianini (2004).

The feature space will be defined by a kernel function

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \tag{3.9}$$

defined on $\mathcal{X} \times \mathcal{X}$ where $\langle \cdot, \cdot \rangle$ is the inner product and $\phi : \mathcal{X} \mapsto \mathcal{H}$ is a (potentially non-linear) function mapping from the input space to a reproducing kernel Hilbert space (RKHS). Thus the kernel function implicitly defines the feature space (or RKHS) by the inner products of the data embedded into $\mathcal{H}$ by ϕ. This direct representation of inner products in $\mathcal{H}$, without explicitly embedding the data points in $\mathcal{H}$ by ϕ is known as the *kernel trick*. The kernel trick also implies that for $N \ll p$ problems, i.e., where the dimensionality p of $\mathcal{X}$ is much larger than the number of observations N, it can be advantageous to work with inner products between observations.

Definition 3.3 (Kernel matrix). The matrix of all inner products between all N observations in $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ is called the kernel matrix

$$\mathbf{K} = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix},$$

also known as the Gram matrix. ▲

Mercer's theorem states which functions are valid as kernel functions (Mercer, 1909, Shawe-Taylor and Cristianini, 2004).

Theorem 3.1 (Mercer's theorem)**.** *A symmetric kernel function* $\kappa(\cdot,\cdot)$ *defined on* $\mathcal{X}\times\mathcal{X}$ *can be defined as an inner product*

$$\kappa(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle$$

for some $\phi:\mathcal{X}\mapsto\mathcal{H}$ *iff*

$$\kappa(\mathbf{x},\mathbf{x}') = \int_{\mathcal{X}\times\mathcal{X}} \kappa(\mathbf{x},\mathbf{x}')f(\mathbf{x})f(\mathbf{x}')d\mathbf{x}dx' \geq 0 \quad \forall L_2(\mathcal{X})$$

or, equivalent that the kernel matrix $\mathbf{K}$ *is positive semidefinite for all sets* $\{\mathbf{x}_i\}_{i=1}^N$.

The symmetry and positive semidefinitenesss of the kernel function ensures that a $\phi:\mathcal{X}\mapsto\mathcal{H}$ exists and the feature space $\mathcal{H}$ is a RKHS. Some popular kernel functions are listed in Table 3.1.

Kernel	Definition $\kappa(\mathbf{x},\mathbf{x}')$	Parameters
Gaussian	$\exp\left(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2\right)$	Scale $\gamma\in\mathbb{R}_+$
Polynomial	$(\langle\mathbf{x},\mathbf{x}'\rangle+c)^d$	Degree $d\in\mathbb{N}_+$, $c\in\mathbb{R}_+$
Sigmoid	$\tanh\left(\gamma\langle\mathbf{x},\mathbf{x}'\rangle+c\right)$	$\gamma,c\in\mathbb{R}_+$
Histogram intersection	$\sum_{i=1}^p \min\{x_i,x_i'\}$	
Chi-square	$1-\sum_{i=1}^p \frac{(x_i-x_i')^2}{\frac{1}{2}(x_i+x_i')}$	

Table 3.1: Some common kernel functions. The Gaussian kernel is the most common, the sigmoid kernel is known from neural networks and the histogram and chi-square kernels are well-suited for histogram feature spaces.

The Gaussian kernel is by far the most popular kernel. In theory it has infinite support, creating infinite dimensional feature spaces. However, the representer theorem (Schölkopf and Smola, 2002) tells us that even though the space $\mathcal{H}$ is infinite-dimensional, the solution to a minimization problem

$$\arg\min_{f\in\mathcal{H}} R((\mathbf{x}_1,y_1,f(\mathbf{x}_1)),\ldots,(\mathbf{x}_N,y_N,f(\mathbf{x}_N))) + g(||f||)\ ,$$

where $R(\cdot)$ is a risk function defined on the set of training samples $(\mathbf{x}_i,y_i)$ with minimizers $f(\mathbf{x}_i)$ and regularizing function $g(\cdot)$, can be written as a linear combination $\mathbf{w}\in\mathbb{R}^N$ of the kernels of each of the N observations

$$f(\mathbf{x}') = \sum_{i=1}^N w_i\kappa(\mathbf{x}_i,\mathbf{x}') \quad \forall i\in\{1,\ldots,N\} \tag{3.10}$$

and thus be maximum N-dimensional. This effectively reduces the search for a minimizer from potentially infinite-dimensional to N-dimensional (Abrahamsen, 2009). The relationship between the Gaussian kernel and the Parzen window estimator is used by Jenssen (2010) to derive kernel entropy component analysis.

Kernels need to be centered in kernel space rather than in the input space. Typically, it is most meaningful to center a set of N_{test} test observations using the set of N training samples. Say that the test observations have been kernelized into the matrix $\mathbf{K} \in \mathbb{R}^{N \times N_{\text{test}}}$. The centered kernel can then be written as

$$\tilde{\mathbf{K}} = \mathbf{K} - \boldsymbol{\mu}_{\text{train}} \mathbf{1}_{N_{\text{test}}}^T - \mathbf{1}_N \boldsymbol{\mu}_{\text{test}}^T + \mu \tag{3.11}$$

where μ_{train} is the N-vector of row-means, μ_{test} the N_{test}-vector of column means and μ the global training mean. In other words, the rows are centered, the columns are centered and the global mean is re-added. In the case of centering a training kernel, i.e., the observations have been kernelized with themselves, computation time can be saved by exploiting the symmetry of $\mathbf{K}$ and realizing that the row and column means are equivalent.

3.3.1 Kernel discriminant analysis

Discriminant analysis refers to the task of classifying previously unseen multivariate observations into a discrete number of classes. This is typically done in a supervised setting, by training a discriminant function on a limited number of observations with known class assignments.

Linear discriminant analysis (LDA) will here be introduced to ease the introduction of kernel discriminant analysis (KDA) in Section 3.3.1.2.

3.3.1.1 Linear and regularized discriminant analysis

The classical Fisher's linear discriminant (also known as canonical discriminant analysis) is a supervised method for classification of multivariate observations (Fisher, 1936).

For a two-class discrimination problem, observe the multivariate data set $\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2$ where $\mathbf{X}_1 = [\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_{\ell_1}^1]^T$ and $\mathbf{X}_2 = [\mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_{\ell_2}^2]^T$ are samples from two different classes with a total of $N = \ell_1 + \ell_2$.

The aim of disciminant analysis is to find the linear combination $\mathbf{Xa}, \mathbf{a} \in \mathbb{R}^p$ that maximizes the variation between classes, while minimizing variation within

classes. This is found by maximizing the objective

$$J(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{S}_B \mathbf{a}}{\mathbf{a}^T \mathbf{S}_W \mathbf{a}} \tag{3.12}$$

where

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{3.13}$$

$$\mathbf{S}_W = \sum_{i=1,2} \sum_{\mathbf{x} \in \mathbf{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{3.14}$$

are the between-class covariance matrix and within-class covariance matrix. $m_i = \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} \mathbf{x}_j^i$ is the class mean vector. This coefficient can be maximized by solving it as a generalized eigenvalue problem.

The mapping of an unseen observation $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$ onto this discriminating direction $z = \mathbf{x}_{\text{new}}^T \mathbf{a}$ yields a scalar z, where negative values classifies the observation into one class and positive values into the other class. For a well defined discriminating manifold, this value can also be interpreted as relative distances to each class.

Oftentimes, it is necessary to regularize the optimization problem by adding a positive definite matrix $\mathbf{\Omega}$ to the within-class covariance matrix (Hastie et al., 1995, Clemmensen et al., 2011). This is especially true for $p > N$ problems, where $\mathbf{\Sigma}_W$ is not necessarily full rank. Often $\mathbf{\Omega} = \lambda \mathbf{I}$ is chosen, where $\mathbf{I}$ is the identity matrix. In that event, the optimization problem in Eq. (3.12) takes the form

$$J(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{S}_B \mathbf{a}}{\mathbf{a}^T (\mathbf{S}_W + \lambda \mathbf{I}) \mathbf{a}} \quad , \lambda \geq 0 \, . \tag{3.15}$$

For $\lambda = 0$ this reduces to the non-regularized discriminant analysis. A two-class classification problem with the LDA solution can be seen in Figure 3.2.

3.3.1.2 Kernel discriminant analysis

Discriminant analysis has been extended to a kernelized version, similar to other multivariate methods, such as principal component analysis (Jolliffe, 2002, Schölkopf et al., 1998). Fisher discriminant analysis with kernels is described well by Mika and Ratsch (1999) and in the context of graph embedding (see Section 3.2) by Cai et al. (2007a).

Here we will jump right into defining the between-class covariance matrix $\mathbf{M}$ in kernel space and the within-class covariance matrix $\mathbf{N}$. These are defined in terms of the kernel function $\kappa(\cdot, \cdot)$. First the mean vector for class i in kernel

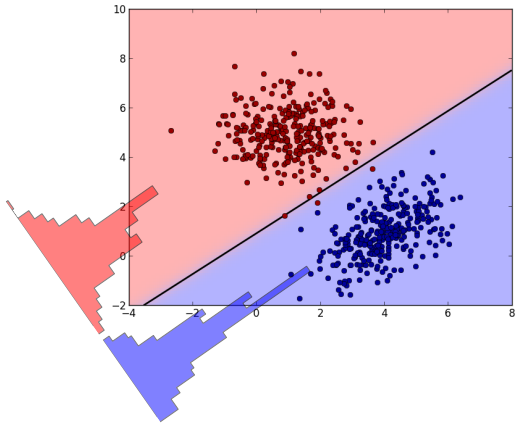


Figure 3.2: Linear discriminant analysis for two classes. The decision boundary separates the domain into two classes. The projected values are shown as histograms and have maximum separation between the two classes.

space is defined as:

$$(\mathbf{m}_i)_j = \frac{1}{\ell_i} \sum_{k=1}^{\ell_i} \kappa(\mathbf{x}_j, \mathbf{x}_k^i) \tag{3.16}$$

and then

$$\mathbf{M} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \tag{3.17}$$

$$\mathbf{N} = \mathbf{K}\mathbf{K}^T - \sum_{i=1,2} \ell_i \mathbf{m}_i \mathbf{m}_i^T \tag{3.18}$$

where

$$\mathbf{K}_{jk} = \kappa(\mathbf{x}_j, \mathbf{x}_k) \tag{3.19}$$

Note that $\mathbf{M}$ and $\mathbf{N}$ are here $N \times N$ matrices.

For the same numerical reasons as before, and the additional need to "capacity control" the feature space, since it can be very non-linear due to the flexibility of the model, it is a must to regularize the within-class covariance.

Similarly to Eq. (3.15) a regularized objective function takes the form as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{M} \mathbf{w}}{\mathbf{w}^T (\mathbf{N} + \lambda \mathbf{I}) \mathbf{w}} \quad \lambda \geq 0. \tag{3.20}$$

This can be solved either as a generalized eigenvalue problem, or, if one is only interested in the direction of the projection vector $\mathbf{w}$, it can be found as $\mathbf{w} = (\mathbf{N} + \lambda\mathbf{I})^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$ (Muller et al., 2001).

It has also been argued that some fraction of the kernel matrix, rather than the identity matrix, could be added for regularization (Nielsen, 2011). This would correspond to penalizing the 2-norm of the projection vector in the original space (Mika et al., 1999). Note once again that $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ is an N-vector, rather than a p-vector as before. Thus a subspace of the feature space is represented in terms of the N observations used to train the discriminant function.

The projection of a new data point using the kernel discriminant function is less trivial than for the linear method. Due to the fact that the kernel method is formulated in terms of individual-similarities (or inner products), the projection of $\mathbf{x}_{\text{new}}$ takes the form

$$z = \sum_{j=1}^{N} w_j \kappa(\mathbf{x}_j, \mathbf{x}_{\text{new}}) \; . \tag{3.21}$$

This can be read as a kernelization of the new observation with each of the training data observations, projected using the discriminating direction $\mathbf{w}$, which of course is a realization of the representer theorem from Eq. (3.10). This implies that the training data set needs to be stored for the testing/classification phase.

For an application of the kernel formulation of discriminant analysis, the reader is referred to papers B and C. This concludes the methodological overview needed to put the scientific contributions into context. Now summaries of the scientific contributions will follow.

Part II

Summary of scientific contributions

Chapter 4

Canonical information analysis

Paper A is motivated by the need for a two-set decomposition method, based on information theoretical measures. Specifically we aim to maximize mutual information between linear combinations of two sets of variables $\mathbf{X} \in \mathbb{R}^{N \times p}$ and $\mathbf{Y} \in \mathbb{R}^{N \times q}$:

$$\arg\max_{\mathbf{a},\mathbf{b}} I(\mathbf{u}, \mathbf{v}) \quad \text{where } \mathbf{u} = \mathbf{X}\mathbf{a}, \mathbf{v} = \mathbf{Y}\mathbf{b} \tag{4.1}$$

where $I(\mathbf{u}, \mathbf{v}) = h(\mathbf{u}) + h(\mathbf{v}) - h(\mathbf{u}, \mathbf{v})$ is the mutual information of the linear combinations. The reason for using mutual information as an optimization criterion is that it accounts for the full probability distribution, rather than, e.g., second order moments only as does canonical correlation analysis (CCA). A motivating toy example is to consider that for two signals, say $y_1(x) = x$ and $y_2(x) = x^2$ the correlation is close to one when $x \in [0, 1]$. However, when $x \in [-1, 1]$ the correlation is exactly zero. In both cases the mutual information is maximum, since the two signals contain the same information. The absence of such method in literature was noted and schematized by De Bie and De Moor (2002) as:

	based on second order statistics	based on mutual information
one signal space	PCA	ICA
more than one signal space	CCA	?
	algorithms use orthogonal projections	algorithms use oblique projetions

Principal components analysis (PCA), independent component analysis (ICA) and CCA were both described in Chapter 3. Only a few attempts towards two- or multi-set information theoretical decomposition have been attempted, though. Noteworthy is the work by Yin (2004), where the same actual optimization problem is being solved by a maximum likelihood estimation with the linear combinations as parameters. However, the method scales poorly with the number of sample points and is as such not well-suited for large-sample problems, such as in image decomposition. Karasuyama and Sugiyama (2012) also propose a solution by directly optimizing the density ratios in the Kullback-Leibler divergence (Eq. (2.18)) by using a basis function representation of the data. However, the presented simulation studies are small-sample and running times are inferior to those of Yin (2004), and thus also not suited for image decomposition problems.

Methodology

The included paper presents "Canonical information analysis" (CIA) as a solution to this problem. The method poses the optimization problem as a general one and focuses on providing a fast approximation of mutual information, given two projection directions. Mutual information estimation can be obtained through marginal and joint entropy estimations. Shwartz et al. (2005) provides an approximate marginal entropy estimator based on quantization and convolution. We have for the purpose of CIA extended this estimator to approximate joint entropy, such that we can obtain an approximate mutual information estimate given the projection directions. This opens the problem up to standard optimization algorithms with many function evaluations, such as simulated annealing or a genetic algorithm, as well as fast, local optimization algorithms, such as the Nelder-Mead downhill simplex algorithm (Nelder and Mead, 1965).

Main results

The main results from the application of canonical information analysis to specific problems are: uncovering of the latent signals in a two-dimensional, two-set toy example is improved, which can be verified visually and numerically. A one-dimensional visualization of eight infrared bands from a weather satellite is visually more pleasing when using CIA to determine the linear combination, compared to a CCA based solution. A change-detection example using two sets of temporally close aerial photos of cars on a highway shows that a potentially more useful difference image is obtained, when jointly determining the linear combinations of the two sets using CIA rather than CCA.

Contributions

Paper A: *Canonical information analysis* holds the following contributions and main results:

- An algorithm for uncovering mutual information maximizing projections of two sets of multivariate data.
- Fast approximate joint entropy estimation.
- Simulation studies for various sample sizes, where it is evident that the proposed method is at least 20 times faster than the method proposed by Yin (2004) for a sample size of 5000, which is a very moderate sample size in image analysis.
- Case studies of remote sensing data of different modalities, illustrating the usefulness of the method.

Several points of future work could be interesting: a kernelized version of CIA, perhaps leveraging the approach by Jenssen (2010), where the Gaussian kernel is used both as the kernel to define the inner products and as the kernel for density estimation to minimize the computational burden. Secondly, application of the method to different data modalities to further emphasize the usefulness of the approach. Finally, the major hurdle to overcome in this approach is the fact that maximization of mutual information is a non-convex optimization problem. Any means to alleviate this would be interesting, especially relaxation of the problem to ensure convergence even for hyperspectral data.

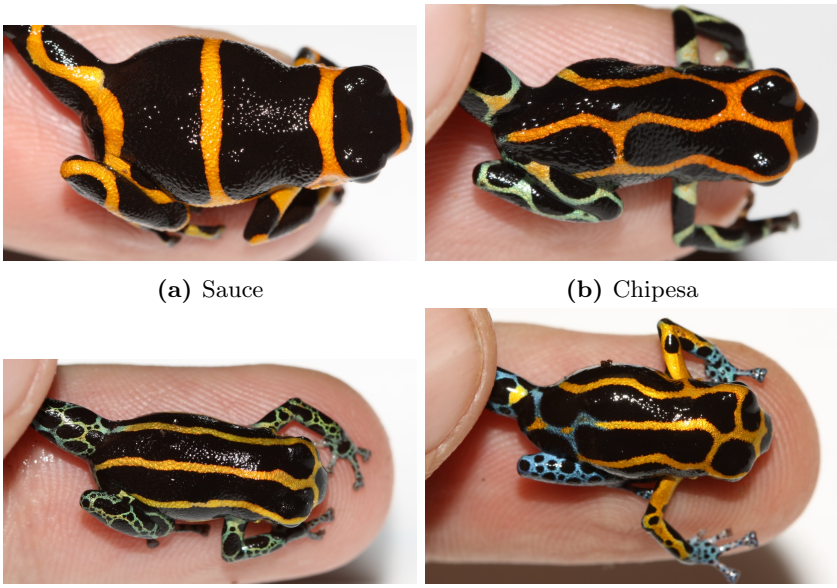
Chapter 5

Quantitative phenotyping of the aposematic frog Ranitomeya imitator

In Papers B, C and D methods are presented for image-based phenotyping of color patterns on poison dart frogs. Paper B details the image analysis methodology, how it is used to quantify a mimicry trait and provides a likelihood model for estimating the number of genes underlying such a trait. In Paper C the methodology is used to quantify the pattern aspects of the phenotype to support the hypothesis that mimicry can be a driver of reproductive isolation. Paper D uses the principles introduced in Paper B to quantify two separate phenotypes from imagery and presents a likelihood model to estimate whether these phenotypes are controlled by the same or separate sets of genes.

The poison dart frog *Ranitomeya imitator* exhibits a complex color pattern as part of its mimicry trait. Different morphs can look very different, despite their genetic similarity (see Figure 5.1). Quantifying this trait is of interest to answer various questions of relevance to evolutionary biologists, and the contributions here are within the field of "quantitative image-based phenotyping". Previously and currently, manual measurement techniques dominate the field of biological quantification of traits. There is an interest in using automated image analysis of field photography to alleviate some common problems associated with manual

measurement techniques: automation is less time consuming for the biologist, the results can be reproduced, and the biases associated with measuring are shifted from subjective, perhaps unkown, biases to the choice of method and parameters.



(a) Sauce **(b)** Chipesa

(c) Achinamisa **(d)** Varadero Forest

Figure 5.1: Individuals of the *R. imitator* species from four different sampling locations in the north-central Peruvian rain forest. Note the remarkable color pattern differences. Also note the variations in pose and illumination.

The mimicry of *R. imitator* has resulted in so-called hybrid zones, where frogs at one of the transect are mimetic with one model species and mimetic with another model species at the other end of this transect. Between these extremes a hybrid zone forms, where intermediate color pattern morphs are found. Quantification of the mimicry trait in such zones can be used to answer the question of, whether mimicry can drive speciation. This is treated in Paper C. Paper B and D exploits that the hybrid zone allows for estimating an admixture proportion for each individual using genetic data (see Appendix J for a data-driven approach for admixture proportion estimation). Such an admixture proportion can be used for biological systems, where it is not convenient to make controlled crosses in the laboratory. Specifically, the admixture proportion is used in connection with the automated quantification of phenotypes to setup likelihood models to answer the questions of 1) "How many genes underly a quantified phenotype?"

and 2) “Are two, separately quantifed, phenotypes controlled by one or separate sets of genes?”.

Methodology

The first step in the presented approach to quantitative image-based phenotyping is to bring the individuals into a common reference frame. Manual annotation of 22 anatomical landmarks was followed by a generalized Procrustes analysis (Gower, 1975), warping the individuals into the average shape. This ensures an anatomically meaningful pixel-to-pixel correspondence which significantly eases further analysis.

A descriptor-based approach is chosen to represent the relevant variation from the images. Specifically, gradient orientation and shape index histograms were used to capture first and second order image information, and a simple binary segmentation was used to separate striped regions from non-striped region. These techniques are described in Section 2.2. To incorporate prior knowledge into this description, it was chosen to spatially average contributions from these descriptors over anatomically meaningful areas; each leg, lower back and head separately. It would be preferable to learn this automatically from the data, but given the fairly small sample size, the complexity of the phenotype and the variability in pose and illumination it was found necessary to explicitly model this.

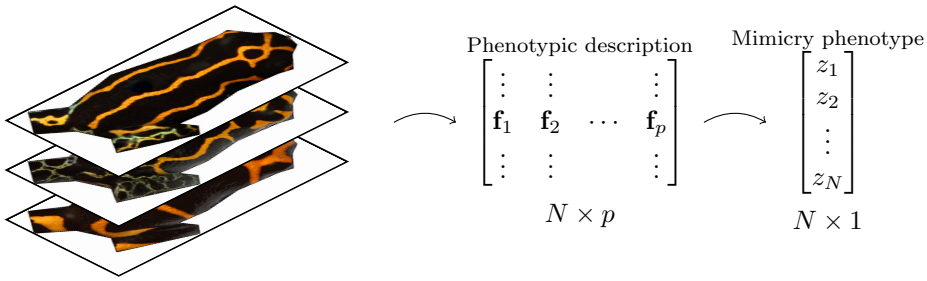
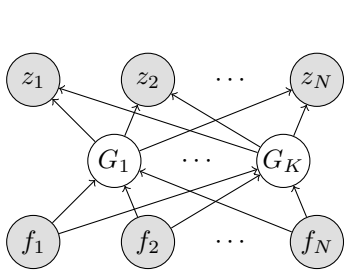
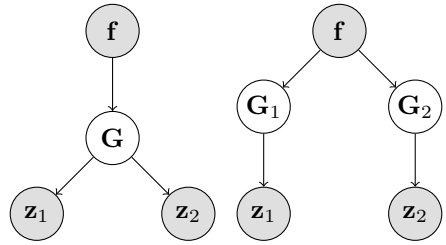


Figure 5.2: Illustrates the process of quantitative image-based phenotyping using image descriptors: A collection of p descriptors are extracted from N images into a $N \times p$ matrix, i.e., a feature matrix describing the phenotype captured by the descriptors. The p dimensions are then reduced to a single dimension, i.e., one scalar per individual, representing the mimicry-aspect of the phenotype. This illustration is also presented in the electronic supplementary material for Paper B.

The extracted phenotypic values are collected in an $N \times p$ data matrix, where N is the number of individuals and p the number of descriptors. This includes members of the two model species. To establish a form of "mimicry index", i.e., how much does an individual resemble each of the model species, the p dimensions need to reduced to a single dimension in a meaningful way. This is illustrated in Figure 5.2. An obvious choice would be to use the first principal axis from a principal components analysis (PCA) of these data (Jolliffe, 2002). However, the first principal axis merely points in the direction of maximum variance, wherefore such a choice will only be meaningful in situations where the mimicry-related aspects of the phenotype constitutes the majority of the variance, which is not necessarily the case. Therefore, in the contributions presented here, variations of Fisher's discriminant analysis (Fisher, 1936) have been used to learn the one-dimensional manifold representing mimicry in this p-dimensional space. The features extracted for the model species were used as training data and the manifold learned as the direction of maximum separation of these two classes. Thus the differences in model species are used to define the manifold of separation. The unlabeled data, i.e., the *R. imitator* individuals, are only used to ensure a smooth manifold. In this context, a smooth manifold is one that does not "collapse" while also ensuring minimum intra-location variance. This is based on the assumption that individuals from the same locations are similar in phenotypic expression.



(a) Model for estimating the number of genotypes K underlying the observed N phenotypes $\{z_i\}_{i=1}^N$ and admixture proportions $\{f_i\}_{i=1}^N$.



(b) Models for testing the hypothesis of whether two phenotypes $\mathbf{z}_1$ and $\mathbf{z}_2$ are dependent on one or two genotypes, given the admixture proportions $\mathbf{f}$.

Figure 5.3: Graphical models for the likelihood models in Papers B and D.

The quantified mimicry-related phenotype and the estimated admixture proportions are key ingredients of the likelihood models presented in Papers B and D. Paper B presents a model useful for estimating the number of genes underlying a quantitative trait in a hybrid zone. This amounts to fitting the proposed model to the observed phenotypes and admixture proportions, with for a varying number of genes and selecting the one with maximum likelihood. Paper D presents a similar model that can be used to identify whether two different quantified

traits are controlled by the same or separate sets of genes. Again, this requires fitting the model to the data and selecting the one with maximum likelihood. This is valid since the models are constructed such that they have the same parameters. Both papers are extensively supported by simulation studies and uncertainty in the estimates are accounted for by a bootstrapping approach. Graphical models of the proposed likelihood models are shown in Figure 5.3.

Main results

Several interesting results are contained in the included papers. These will be summarized for each paper here:

Paper B: *Number of genes controlling a quantitative trait in a hybrid zone of the aposematic frog Ranitomeya imitator* This paper analyzes a hybrid zone containing 317 individuals, stretching from Sauce along the Huallaga river to Micaela Bastidas. In this zone, the two model species are *R. summersi* and *R. variabilis*. Using a 60-dimensional description of pattern, which is reduced to a scalar mimicry-related phenotypic quantity per individual, it is found most likely that the mimicry is controlled by one, two or at most three genes of major effect.

Paper C: *Reproductive isolation related to mimetic divergence in the poison frog Ranitomeya imitator* This paper contains several analyses supporting the hypothesis of reproductive isolation. This includes mate choice experiments, bioacoustical analysis, landscape genetics, and color pattern analysis, the first three being unrelated to the contributions of this thesis. The employed pattern analysis methodology is similar to that of Paper B, but is here used to analyze another transition zone with *R. fantastica* and *R. variabilis* being the two model species. For this paper, it was relevant to quantify the mimicry-related phenotype separately for the legs and the body of the frog as divergent selection may act differently on different phenotypes. Based on spectroscopic measurements a similar measure for arm, leg, head and body color was derived. Collectively, these quantifications serve as the foundation for statistical analyses showing the presence of a narrow phenotypic transition zone. Combined with neutral genetic divergence and assortative mating, shown by the other analyses, this supports that mimicry-driven speciation is in an early stage for this vertebrate system.

Paper D: *Identifying pleiotropic control of adaptive phenotypes* This paper uses the same hybrid zone as in Paper B and sets up four specific

cases to analyze, namely whether the phenotype pairs of dorsal saggital vs. transversal stripes, dorsal pattern vs. dorsal coloration, dorsal pattern vs. leg pattern and dorsal coloration vs. leg coloration are controlled by one or separate sets of genes. The first case is included as a very simple example of a quantifiable phenotype; the remaining phenotypes are quantified as in Paper B. The proposed likelihood model is fitted to these pairs of phenotypes and it is found for the first three cases that the pairs are most likely controlled by the same set of genes. There is no evidence for this in the case of dorsal vs. leg coloration.

To make this result conceivable, a reaction-diffusion model for pattern formation is derived. This model is capable of generating patterns that loosely resemble those found in the transition zone, changing only a single parameter to go from saggital stripes to transversal stripes and dots for an intermediate value. See Appendix K for more detail on reaction-diffusion models.

Contributions

The main contributions from the work related to quantitative phenotyping of mimicking frogs in a hybrid zone are:

- Automated extraction of useful features related to the pattern phenotype from field imagery, i.e., varying light, pose and camera position.
- Estimation of a "mimicry index" from these extracted features using manifold learning.
- Derivation of likelihood models for testing two different hypotheses relevant to evolutionary biologists: how many genes underly a quantitative trait, and are two separately quantified phenotypes controlled by the same or separate sets of genes?
- Practical implementations and simulations documenting the precision and accuracy of these methods.
- Reaction-diffusion models and simulations for pattern generation.
- Estimation of admixture proportions from microsatellite data using kernel discriminant analysis and a kernel employing a genetic distance measure.
- Application of the developed methodology to three different hybrid zones with a total of 588 individuals.

Interesting points to continue work in this direction on would be a generative model for the pattern description. This could be modelled in various ways, e.g., in a patch-based deep-learning architecture, where the low-dimensional representation is ensured to be able to generate the observed images. Another point of potential improvement would be the reduction from a multivariate phenotypic representation to a one-dimensional mimicry-related manifold. The determination of the manifold could be attempted using variations of, e.g., logistic regression or, perhaps more interesting, an approach directly modelling group membership probabilities, such as (sparse) mixture discriminant analysis (Hastie and Tibshirani, 1996, Clemmensen et al., 2011).

Chapter 6

Structure identification in graphene

In Paper F: *Pattern recognition approach to quantify the atomic structure of graphene* we motivate the usefulness of an automatic method for determining the structure from low contrast HRTEM images. In Paper E: *Structure identification in high-resolution transmission electron microscopy images: an example on graphene* we describe the methodology constituting the pipeline.

Graphene is a two-dimensional material as it is only one atom thick. The layer of atoms is arranged in a hexagonal structure, a honey-comb lattice. In an unaltered, pristine, graphene sheet, this structure is completely regular. However, the really interesting semi-conducting properties of graphene only emerge when the sheet is altered by puncturing periodic holes. The atomic structure surrounding such holes are less stable, i.e., they change continuously and the sheet may even bend or buckle due to the stress. While simple methods, such as extrema detection followed by a triangulation, could be used for determining the structure of a pristine graphene sample, the altered graphene poses different challenges. In Appendix H an image registration inspired approach is described, which was found sufficient for a large part of the samples (Kling et al., 2013). However, the method was slow and not in a framework capable of handling local adjustments. Instead, the method was matured by employing a Markov random field model inspired by Hartelius and Carstensen (2003). We distinguish

ourselves from the work by Hartelius and Carstensen (2003) in that we do not fit a prior known grid to an image. Rather, a set of points are detected in the image that are known to originate from a known grid *structure*. This has the conceptual advantage of not having to specify the grid instance in advance, and the computational tractability of reducing the problem to a set of well initialized points.

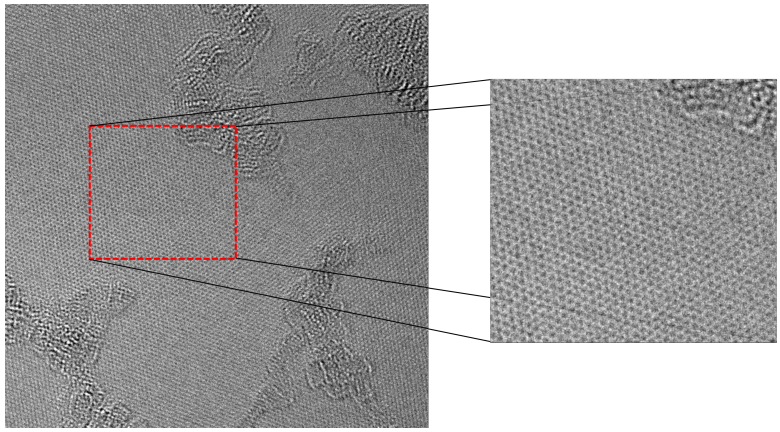


Figure 6.1: Example of an imaged graphene sheet and an excerpt. The full image is 2048×2048 pixels (24.48 nm $\times$ 24.48 nm). Note that the dark spots are the hexagonal centers and thus form the dual triangular lattice when connected. The irregular areas are amorphous graphene, a side-effect from the manufacturing process.

The presented pipeline for determining the atomic structure consists of four steps:

1. Determine the approximate lattice scale: This is done by leveraging the 2D Fourier response of the lattice in which the periodicity can be determined and converted to lattice properties.

2. Local minima detection: The local minima found – and what will be referred to as the sites – are the centers of the hexagons in the lattice, seen as dark spots in Figure 6.1. The centers of the hexagonal lattice form the dual lattice, which is triangular. The scale found in the first step is used to guide the minima detection.

3. The neighborhood structure of the detected centers needs to be inferred. For MRFs this is assumed known in advance, which is not the case here since the lattice is initialized in the observed image rather than laid on top of it. An iterative heuristic is used to infer this neighborhood structure:

First an initial neighborhood is constructed using a Delaunay triangulation, secondly improbable connections are removed based on knowing that the mesh should consist of approximately equilateral triangles, where the side length is approximately known from the Fourier analysis. Thirdly, points with very little observation power (weak minima) are removed. The process is repeated until the mesh does not change anymore. In the second step, triangles and points are removed by setting up an Ising model inducing spatial homogeneity. While this is a heuristic for determining neighborhood structure, it was found to work robustly for all cases treated.

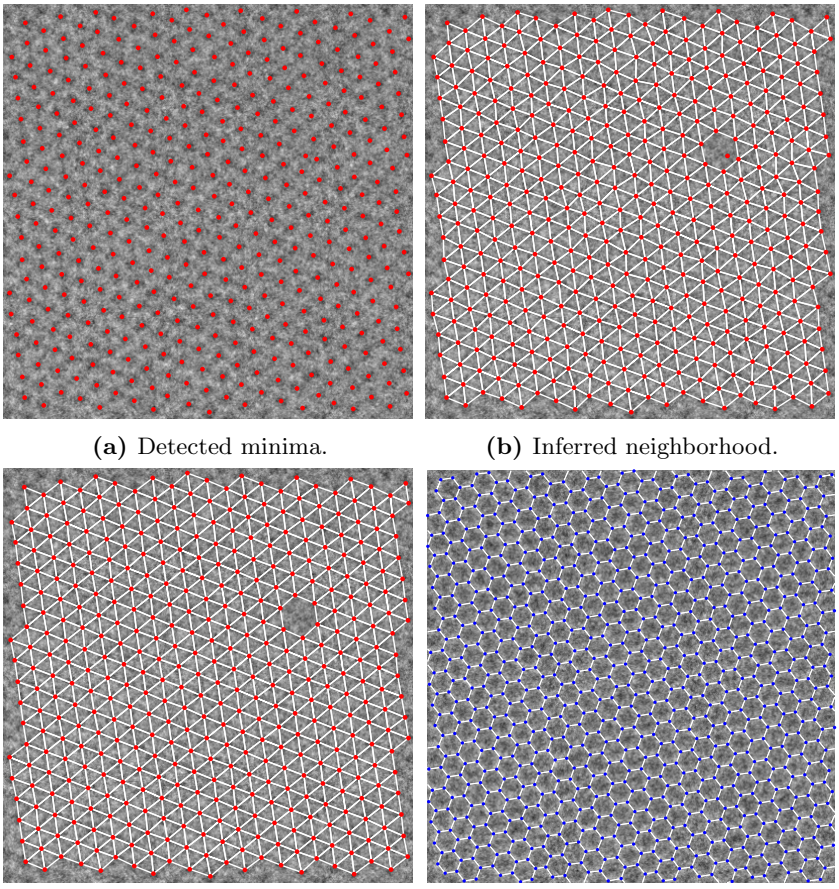
4. Having determined a neighborhood structure, a posterior model combining the geometric prior and the observation model (as described in Section 2.3.4) is formulated. The minimum energy configuration is found by generating moves in the random walk using the Metropolis spin-flip algorithm and simulated annealing.

The last three of these steps are illustrated in Figure 6.2, where the final honey comb lattice is also shown.

The microscopic structure is of interest for material scientists, e.g., carbon-carbon bond lengths or other properties that can be inferred from the estimated lattice. Therefore it is of interest which carbon-carbon bond lengths are unexpectedly short and provide a visualization of this. We leverage the framework of false discovery rate large-scale simultaneous hypothesis (FDR-LSSHT) testing by Efron (2004) to provide a statistically sound interpretation of this. FDR-LSSHT allows for multiple testing of a large number of hypotheses, while setting a maximum for the proportion of false positive tests and thus alleviates the conservatism of normal multiple comparison methods, e.g., Bonferroni adjustment, which in high-dimensional problems tend to lead to too few significant variables.

Main results

The main results from this work is that we can extract parameters of the microscopic structure in graphene sheets produced under various conditions. In the two papers included, we show that the distributions of estimated carbon-carbon bond lengths differ between a pristine graphene sheet and an altered graphene sheet with an induced hole. For easier comparison, we here re-present the histogram comparison, representing the main result from the two papers in Figure 6.3, as cumulative distribution functions as a supplement to the histograms presented in the papers. The shown distributions represent ten exposures of the



(a) Detected minima. **(b)** Inferred neighborhood.

(c) Adjusted positions. **(d)** Inferred graphene structure.

Figure 6.2: The steps of the algorithm illustrated on a small excerpt of the graphene sheet in Figure 6.1. Note how one of the detected minima is not connected due to its unexpected position in the grid. In this case, carbon atoms are still placed around the missing hexagon center, due to its neighboring centers.

pristine sample (red) and twenty exposures of the altered sample (blue); all lattices are fitted using the same set of parameters for the algorithm. The bond length distributions for the altered sample show a heavy left tail indicating an abundance of shorter bond lengths, which in turn can be interpreted by material scientists as either bending/buckling of the sheet or an actual shortening of bond lengths. Together with visualizations of the FDR-LSSHT interpretation presented in the papers, this provides a statistically meaningful foundation for further hypothesis confirmation or generation. Note that the two distributions compared in Paper F are two of the individual exposures shown here and in Paper E.

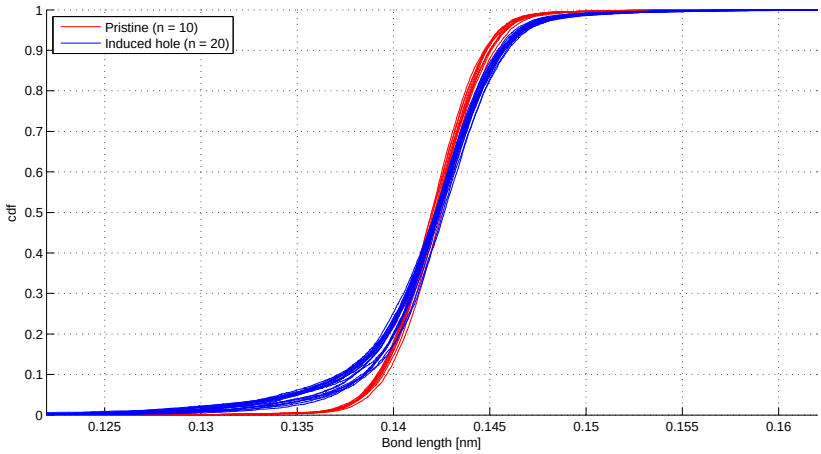


Figure 6.3: Bond length distributions estimated for ten exposures of a pristine graphene sample (red) and twenty exposures of an altered graphene sample (blue). The distributions are shown as cumulative distribution functions. The distributions are based on approximately 16100 estimated bond lenghts for the pristine sample and approximately 11800 bond lengths for the altered sample.

Contributions

The main contributions of the work related to structure identifcation in graphene are:

- Automated extraction of graphene sample parameters from frequency analysis of an HRTEM image.

- Derivation of the needed mathematical models for fitting a triangular grid to the main image evidence, and inferring the atomic structure from this fit.
- Practical implementations of these methods with proofs of concepts.
- Extraction of relevant parameters, visualization and statistical treatment of these using the framework of LSSHT.
- Application of these methods to multiple exposures of two graphene samples.

Two specific extensions of this method would be interesting. Determining the neigborhood structure by the alternating graph cut approach could be replaced by a soft weighting with the probability of two nodes being neighbors. This could make the optimization of the nodes less deterministic and perhaps provide a richer result for further analysis. However, it would require a maximum aposteriori estimate of the final grid structure, that in fact honors the geometry, to provide a visualization of carbon atoms placement. The second extension would be an expanding grid, i.e., an expansion of the parameter space while optimizing the node placements. This is less trivial in that it requires a mechanism for expanding the grid that is geometrically meaningful and satisifies detail-balance. The latter is non-trivial, but a starting point would be the literature on reversible jumps, introduced by Green (1995).

Chapter 7

Classification of polarimetric SAR data using dictionary learning

In Paper G: *Classification of Polarimetric SAR Data Using Dictionary Learning* an image patch-based classification is used for a multi-class crop classification problem.

Polarimetric synthetic aperture data (SAR) provides rich reflectance data in that it transmits horizontally (H) and vertically (V) signals and receives them horizontally and vertically as well. This provides the four polarizations HH, HV, VH and VV. Due to symmetry these signales are collected in a three-vector $\mathbf{k} = [S_{HH}, S_{HV}, S_{VV}]^T$ and a 3×3 multilook covariance matrix is set up as $\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{k}(i)\mathbf{k}(i)^T$ where n is the number of looks (Skriver, 2012). "Multilook" simply means spatial averaging and could also be referred to as gridding. Previous methods have relied on statistical assumptions or geometric knowledge of the SAR signal, (see e.g. Cloude and Pottier, 1997, Hoekman and Vissers, 2003).

The presented paper poses the crop classification problem as a supervised classification problem and leverages the patch-based discriminative clustering method by Dahl and Larsen (2011). This involves expanding the covariance matrix in

each pixel to a nine-vector given by three diagonal elements of $\mathbf{Z}$ plus real and imaginary parts of the three unique off-diagonal elements. The feature space is extended spatially, due to patches, and temporally, due to multiple acquisition points. The discriminative clustering is therefore carried out in a feature space with a dimensionality dependent on the polarization mode, the patch size and the number of temporal acquisitions. Training of the model is done by dividing the image into a training region and a testing region; this division makes only approximately 32% of the image available as training data to create a challenging scenario for the algorithm.

Main results

The obtained classification is evaluated in terms of classification error percentage for various choices of image patch size. Single (HH and VV separately), dual (HHVV) and full polarization modes are used as input data. We compare the classification results to a maximum likelihood classifier described by Skriver (2012) and find that the patch-based classification is superior in all cases in terms of misclassification rate. The maximum likelihood classifier, however, shows comparable performance for fully polarimetric data when including three or four acquisition points.

Contributions

The main contributions from this work are:

- Demonstrating the usefulness of clustering algorithms in the context of polarimetric SAR data.
- Extension of existing methodology for patch-based clustering to incorporate time series information.
- A scheme for spatial division of data into training and test sets allowing for cross-validatied error estimates.
- Comparison of the proposed patch-based classifier and an existing pixel-based maximum likelihood classifier.

An obvious improvement of this clustering based classification technique would be to leverage the known statistical properties of multilook polarimetric SAR

data studied by Conradsen et al. (2003). This could readily be done by using the derived test statistic for equality of two complex Wishart matrices as the distance measure used for clustering the observations.

Conclusions

This thesis provides an overview of methodology relevant to quantitatively describe image content in terms of intensity, texture and geometry, and how to choose an appropriate representation of said contents. The methodological description and approach is general, while the application of these methods is adapted specifically to the problem at hand.

The main problems treated in this thesis have been the cases of quantifying the atomic structure of graphene from high-resolution transmission electron microscopy images (HRTEM), quantitative phenotyping of the aposematic frog *Ranitomeya imitator* from field imagery, identification of crops from polarimetric synthetic aperture radar (SAR) data, and mutual information based two-set decomposition of multivariate imagery.

In the case of determining the atomic structure of graphene, the main contributions have been: devlopment of a pipeline of methods capable of estimating the hexagonal grid structure of carbon atoms from a single HRTEM image, extraction of relevant parameters from a fitted grid, and visualizing extracted parameters in a statistically meaningful way. The pipeline leverages classical image analysis, two-dimensional Fourier analysis and probabilistic graphical models to solve the problem robustly, while honoring prior available information.

Quantitative and reproducible description of the color pattern polymorphism of *R. imitator* is a relevant problem for evolutionary biologists. The main contributions within this topic are automated extraction of features relevant to a specific phenotype from field imagery, quantifying the aspects of the extracted features relevant for mimicry, and development of likelihood models capable of answering fundamental questions in evolutionary biology. Collectively, these

contributions also serve the purpose of making image analysis available to biologists, by illustrating the usefulness of using methods that are reproducible. Further, this approach move the bias from the subtle subjectivity of the biologist to the more apparent parameter choices relevant for the method.

A compressed representation of the polarimetric SAR imagery is shown useful for classification of crops. While compressed sensing is a well-known concept in the field of image analysis, it had not previously been used in a polarimetric SAR classification context. Through a cross validation scheme a classification algorithm is trained to classify image patches into six different classes. The approach is shown to be superior to the existing state-of-the-art maximum likelihood, pixel-based, classification method.

Mutual information based two-set decomposition of multivariate imagery is a case of pure methodological development. The contribution from this work is that it fills a gap in the availability of two-set decomposition methods, namely that of an information theoretical approach maximizing mutual information. While previous attempts exist in literature, these are not well-suited for large-sample problems such as images. Through fast convolution-based approximations of the entropy estimates, canonical information analysis is presented as a viable solution to the problem.

The objectives of the thesis have been met through development of general, flexible, methods for describing image intensity, texture and geometry in the input space, and representing this information in an appropriate feature space. Through collaboration- and problem-driven development of the necessary methodology, interesting questions have been answered in research domains as diverse as evolutionary biology and materials science. This thesis makes statistical image analysis available to fellow researchers with domain specific problems, and provides new methodology relevant for the field itself.

Included papers

Paper A

Canonical information analysis

Canonical Information Analysis

Jacob Schack Vestergaard[a,*], Allan Aasbjerg Nielsen[a]

[a]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Asmussens Alle, Building 324, DK-2800 Lyngby, Denmark.*

Abstract

Canonical correlation analysis is an established multivariate statistical method in which correlation between linear combinations of multivariate sets of variables is maximized. In canonical information analysis introduced here, linear correlation as a measure of association between variables is replaced by the information theoretical, entropy based measure mutual information, which is a much more general measure of association. We make canonical information analysis feasible for large sample problems, including for example multispectral images, due to the use of a fast kernel density estimator for entropy estimation. Canonical information analysis is applied successfully to 1) simple simulated data to illustrate the basic idea and evaluate performance, 2) fusion of weather radar and optical geostationary satellite data in a situation with heavy precipitation, and 3) change detection in optical airborne data. The simulation study shows that canonical information analysis is as accurate as and much faster than algorithms presented in previous work, especially for large sample sizes.
URL: `http://www.imm.dtu.dk/pubdb/p.php?6270`

Keywords: Information theory, probability density function estimation, Parzen windows, entropy, mutual information maximization, canonical mutual information analysis, CIA, approximate entropy

1. Introduction

In canonical correlation analysis (CCA) first published by Hotelling in 1936 (Hotelling, 1936) linear combinations $U = \mathbf{a}^T\mathbf{X}$ and $V = \mathbf{b}^T\mathbf{Y}$ of two sets of stochastic variables, k-dimensional $\mathbf{X}$ and ℓ-dimensional $\mathbf{Y}$, which maximize correlation between U and V are found. Correlation considers second order statistics of the involved variables only and as such it is ideal for Gaussian data. In this paper we investigate replacement of correlation with mutual information (Hyvärinen et al., 2001; Mackay, 2003; Bishop, 2007; Canty, 2010) which is a more general, information theoretical, entropy based measure of association between variables. Entropy and mutual information (MI) depend on the actual probability density functions of the involved variables and thus on higher order statistics. The resulting method is termed canonical mutual information analysis, or in short canonical information analysis (CIA).

Since multi-source data, which is typically of different genesis, often follow very different (non-Gaussian) distributions, the application of MI facilitates analysis of such data. In one of our examples we apply the method to a joint analysis of radar and optical data (which follow very different distributions thus rendering CCA non-optimal). Other areas where the method could potentially be very useful include data of different modalities, for example SAR, LiDAR, optical and medical data. In general, this type of analysis has a strong potential for application in data fusion and other fields of data integration, see also (Ehlers, 1991; Pohl and Van Genderen, 1998; Conese and Maselli, 1993).

Mutual information as a measure of association has previously proven useful in the context of image registration. Studholme et al. (1999) proposed a normalized variant of MI for registration of medical images, which Suri and Reinartz (2010) employ for automatic registration of SAR and optical images. For the purpose of change detection, Erten et al. (2012) derive an analytical expression for the mutual information between temporal multichannel SAR images.

*Corresponding author
Email addresses: `jsve@dtu.dk` (Jacob Schack Vestergaard), `alan@dtu.dk` (Allan Aasbjerg Nielsen)
URL: `http://www.compute.dtu.dk/~jsve` (Jacob Schack Vestergaard)

Other dependence measures have been considered in the literature, such as kernel canonical correlation analysis (kCCA) (Lai and Fyfe, 2000; Bach and Jordan, 2002). However, while kernel methods do indeed provide an implicit nonlinear transformation of the data maximizing some dependence measure, they do not possess the same qualities as linear methods in terms of interpretation. Specifically, a linear method, such as CIA, finds the actual functional relation between the original variables, where a kernel method, such as kCCA, would find a hidden/intrinsic transformation which makes the relation between CVs linear. This property of the linear solution immediately eases interpretation of the result.

The idea of maximizing MI between two sets of variables is mentioned by de Bie and de Moor (2002). However, the authors only propose solutions to this problem based on independent component analysis in the individual spaces of the variables and they do not provide a truly canonical approach. Yin (2004) and Karasuyama and Sugiyama (2012) solve the problem of maximizing MI of linear combinations of variables in a manner which makes its application to small sample problems feasible. In practical terms the solutions offered are not applicable to large sample problems including for example image data. Our fast grid-based entropy estimator (Section 5) facilitates the use of CIA to large sample problems. Both Yin (2004) and Karasuyama and Sugiyama (2012) request orthogonality between solutions (as in CCA), whereas we allow for oblique solutions (Section 2) via a structure removal procedure inspired by Friedman's Projection Pursuit (Friedman, 1987). The well known difficulties in estimating and optimizing entropy measures, will be addressed in Sections 4, 5 and 6.

Below, Section 2 describes the concept of canonical information analysis and motivates the following sections. Section 3 describes the information theoretical concepts entropy of a univariate stochastic variable, joint entropy of two stochastic variables, relative entropy, and mutual information. Section 4 briefly describes the estimation of one- and two-dimensional probability density functions, Section 5 describes approximate entropy estimation, and Section 6 describes the maximization of mutual information of two linear combinations of stochastic variables. Section 7 gives 1) a simple, illustrative toy example, 2) a case study with weather radar data and optical data from a meteorological satellite, and 3) a case with change detection in optical airborne data. Section 8 concludes. An appendix is included, motivating some of the implementation choices made. Supplementary material is provided with additional simulation studies and results from the two case studies plus an extra application of CIA for change detection.

2. Canonical Information Analysis

Inspired by canonical correlation analysis (Hotelling, 1936) we propose a method for maximizing mutual information between the linear combinations $U = \mathbf{a}^T\mathbf{X}$ and $V = \mathbf{b}^T\mathbf{Y}$ of two sets of stochastic variables, k-dimensional $\mathbf{X}$ and ℓ-dimensional $\mathbf{Y}$.

The goal of CIA can be stated as

$$\mathbf{a}^\star,\mathbf{b}^\star = \arg\max_{\mathbf{a},\mathbf{b}} I(U,V) \tag{1}$$

where $I(U,V)$ is the mutual information between the two linear combinations U and V which can be defined as

$$I(U,V) \quad = \quad h(U) + h(V) - h(U,V) \tag{2}$$

where $h(U)$ and $h(V)$ are the marginal entropies and $h(U,V)$ the joint entropy. This will be detailed further in Sections 3, 4 and 5.

Maximization of mutual information is known to be a non-convex optimization problem (Modersitzki, 2004; Haber and Modersitzki, 2007) wherefore we have conducted experiments with local as well as global optimization methods, see Section 6. The inherent lack of certainty of finding a global optimum will be elucidated by application of the method to different real world multispectral decomposition problems, see Section 7.

In canonical correlation analysis k and ℓ linear combinations (components) are determined with the criterion that the i'th component maximizes correlation between U and V while being orthogonal to the first $i-1$ components. Friedman (1987) introduced in projection pursuit 'structure removal' as the solution to avoid re-finding a previously found direction in space. Structure removal works by histogram equalization of the projected data to a Gaussian distribution and transforming back to the original space. In CIA we choose to adopt this principle of structure removal with the modification that the projected data U and V are substituted with uniformly distributed white noise. This

modification is necessary since, in contrast to projection pursuit, CIA does not maximize non-Gaussianity of one projection, but rather it maximizes statistical dependence between two projections. This structure removal replaces the orthogonality requested by Yin (2004) and Karasuyama and Sugiyama (2012).

3. Basic Information Theory

In 1948 Shannon (Shannon, 1948) published his now classical work on information theory. Below, we describe the information theoretical concepts entropy and mutual information for discrete and continuous stochastic variables, see also (Hyvärinen et al., 2001; Mackay, 2003; Bishop, 2007; Canty, 2010).

3.1. Discrete variables

Consider a discrete stochastic variable X with probability density function (pdf) $p(X = x_i)$, $i = 1,\ldots,N$. The information content is defined as $-\ln(p(X = x_i))$. The expectation $H(X)$ of the information content is termed the entropy of the stochastic variable X

$$H(X) = -\sum_{i=1}^{N} p(X = x_i)\ln(p(X = x_i)). \tag{3}$$

For the joint entropy of two discrete stochastic variables X and Y we get

$$H(X,Y) = -\sum_{i,j} p(X = x_i, Y = y_j)\ln(p(X = x_i, Y = y_j)). \tag{4}$$

3.2. Continuous variables

Probability density functions, information content and entropy may be defined for continuous variables also. This is necessary to represent linear combinations of sampled data. In this case the entropy

$$h(X) = -\int p(x)\ln(p(x))\mathrm{d}x \tag{5}$$

is termed differential entropy. Since $p(x)$ here may be greater than 1, $h(X)$ in the continuous case may be negative (or infinite).

Empirical entropy $\hat{h}(X)$ is an estimator of $h(X)$ in (5). The estimator is defined as

$$\hat{h}(X) = -\frac{1}{N}\sum_{i=1}^{N}\ln(p(X = x_i)) \tag{6}$$

and as such it is defined over a finite sample $\{x_i\}_{i=1}^{N}$ of X, where N is the number of samples. As opposed to (3) and (4) this estimator is not based on any binning of the data.

Empirical entropy has previously proven useful for manipulating entropy measures (Viola, 1995). We have experienced this experimentally (not shown here) and find this estimator useful for canonical information analysis.

The extent to which two continuous stochastic variables X and Y are not independent, which is a measure of their mutual information content, may be expressed as the relative entropy or the Kullback-Leibler divergence between the two-dimensional pdf $p(x,y)$ and the product of the one-dimensional marginal pdfs $p(x)p(y)$, i.e.,

$$D_{KL}(p(x,y), p(x)p(y)) = \int\int p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y. \tag{7}$$

This sum defines the mutual information $I(X,Y) = D_{KL}(p(x,y), p(x)p(y))$ of the stochastic variables X and Y. Mutual information equals the sum of the two marginal entropies minus the joint entropy

$$I(X,Y) = h(x) + h(y) - h(x,y). \tag{8}$$

Unlike the general Kullback-Leibler divergence this measure is symmetric. Mutual information is always nonnegative, it is zero for independent stochastic variables only.

We need to estimate marginal as well as joint pdfs to obtain the mutual information estimate in (8). Karasuyama and Sugiyama (2012) estimate the ratio in (7) directly. We employ kernel density estimation, which uses N data samples to estimate these pdfs. Mutual information is subsequently estimated using the same N data points. This is possible in practice only due to our very fast estimation of pdfs which will be described in Section 5. Note, that this is in contrast to Viola (1997) where the sample is divided into smaller portions in order to lessen the computational burden and to Yin (2004) where an explicit estimation is used that does not scale well to image analysis problems and other large sample problems.

4. Density Estimation

The histogram is a simple non-parametric density estimator. However, the estimated histogram is not smooth and it depends on the end points of bins and the width of bins. By using kernel density estimators (Rosenblatt, 1956; Parzen, 1962; Silverman, 1986) where we center a kernel on each observation, we may obtain smoother histograms that do not depend on bin end points. The kernel density estimator (Parzen windows estimator) for the pdf of X at value t is

$$\hat{p}(X = t|\mathbf{x}) = \frac{1}{N\sigma} \sum_{i=1}^{N} \varphi\left(\frac{t - x_i}{\sigma}\right) \tag{9}$$

where $\mathbf{x} = \{x_i\}_1^N$ is a vector of realizations of X, $\varphi(z)$ is the kernel and σ a smoothing parameter referred to as the bandwidth. Often we choose the Gaussian kernel

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) . \tag{10}$$

The width of the Gaussian, i.e., the standard deviation is thus equivalent to the bandwidth σ.

The kernel density estimator assumes continuous distributions, thus we estimate continuous variants of the information theoretic measures mentioned in Section 1. Since only two one-dimensional projections of the data are considered, the known problems with kernel density estimators in higher dimensions (Beirlant et al., 1997; Kraskov et al., 2004) are found to be negligible for canonical information analysis.

In two dimensions the bivariate Gaussian is often chosen to have a diagonal covariance matrix leaving two parameters to be estimated, namely the bandwidth in each direction. Estimation of the bandwidth is an example of the bias-variance trade-off: a too narrow kernel causes too large variation in the density estimate and a too wide kernel oversmooths the estimated distribution (Jones and Marron, 1996).

Here we use a data-driven bandwidth selection method based on the maximal smoothing principle (Terrell, 1990). This method is known to be conservative (oversmoothing) by nature (Jones and Marron, 1996; Terrell, 1990), but this is outweighed by fulfilling two – in this context – more important properties: the bandwidth estimate is stable, i.e., it varies smoothly for small changes in projection direction of the data. Experiments (see Appendix A) have shown that this is not the case for, e.g., neither the linear diffusion process based method by Botev et al. (2010) nor for Sheather-Jones (Sheather and Jones, 1991). The second property is computational speed, where it outperforms the commonly preferred "solve-the-equation plug-in" method (Sheather and Jones, 1991). Speed is of practical importance as the density estimation will be part of calculating the objective value for a non-convex optimization problem, wherefore the bandwidth will be estimated repeatedly. This is especially true for large problems, e.g., image processing.

5. Approximate Entropy Estimation

Estimation of marginal and joint entropies is the main bottleneck in maximization of mutual information. Parzen window density estimation, in the explicit form presented above, has previously been used for this purpose, see e.g. Yin (2004). However, since it is based on pairwise distances, it has a computational complexity in the order of $O(N^2)$. Shwartz et al. (2005) proposed a fast approximate marginal (1D) entropy estimator with a complexity in the order of

$O(N \log N)$. For the purpose of canonical information analysis we generalize this approximate entropy estimator to joint entropy (2D). This is described below and illustrated in Figure 1.

Approximate entropy estimation is a convolution based modification of Parzen window density estimation. Convolution of the samples with the kernel in (10) is equivalent to the density estimation in (9). Convolutions can run in the order of $O(N \log N)$ on a regular grid. The estimation procedure therefore 1) quantizes the irregular samples to a regular grid, 2) convolves with a Gaussian kernel on this grid, and 3) interpolates back onto the samples' original positions to get an estimate of the empirical entropy in (6).

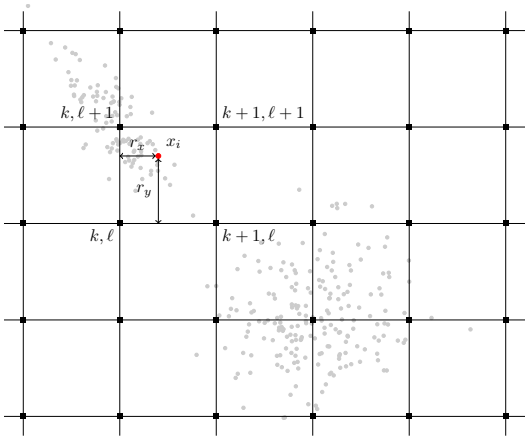


Figure 1: **Quantization** Illustration of bilinear quantization of samples to a regular grid to enable fast approximate joint entropy estimation. The gray dots are examples of irregular samples and the red dot is used to exemplify the bilinear weights. The black rectangles indicate the bins and the indices of the four bins influenced by the red dot are shown.

Quantization requires choosing a discretization, i.e., a number of bins B^2 and a domain $[x_a, x_b] \times [y_a, y_b]$ over which to discretize. The (m,n)'th bin in this regular grid is positioned at $(x,y)_{m,n} = (x_a + m\Delta x, x_b + n\Delta y)$ where $m,n \in \{0, \ldots, B-1\}, \Delta x = \frac{x_b - x_a}{B-1}$ and $\Delta x = \frac{y_b - y_a}{B-1}$. The i'th sample point falls into a cell spanned by the four bin centers with indices $(m_i, n_i), (m_i + 1, n_i), (m_i, n_i + 1), (m_i + 1, n_i + 1)$ where $m_i = \text{fl}\left[\frac{x_i - x_a}{\Delta x}\right]$ and $n_i = \text{fl}\left[\frac{y_i - y_a}{\Delta y}\right]$ and fl$[\cdot]$ is the floor operation. The weights for each of these four bin centers are given by a bilinear interpolation scheme:

$$\begin{aligned} w_i(m_i,n_i) &= (1 - r_x)(1 - r_y) \\ w_i(m_i + 1,n_i) &= r_x(1 - r_y) \\ w_i(m_i,n_i + 1) &= (1 - r_x)r_y \\ w_i(m_i + 1,n_i + 1) &= r_x r_y \end{aligned}$$

where $r_x = \frac{x_i - x_a}{\Delta x} - m_i$ and $r_y = \frac{y_i - y_a}{\Delta y} - n_i$, i.e., the fraction removed by the floor operation. The quantized value $Q_{m,n}$ at a given bin is thus a weighted count of samples in the proximity of the bin. The quantization is collected in a $B \times B$ image-like matrix $\mathbf{Q}$. This bilinear weighting is the 2D analogue of the linear weighting suggested by Shwartz et al. (2005).

Convolution of the quantized signal on the regular grid with the kernel φ from (10)

$$\hat{\mathbf{Q}} = \varphi * \mathbf{Q}$$

can be performed in the order of $O(B^2 \log B^2)$, i.e., dependent on the number of bins rather than the number of samples. The resulting (m,n)'th element of $\hat{\mathbf{Q}}$ is an estimate of the density at the (m,n)'th bin. Distributing this estimate back onto the original sample positions is done using the same weights as earlier, such that

$$\hat{p}(x_i) = Q_{m_i,n_i} w_i(m_i,n_i) + Q_{m_i+1,n_i} w_i(m_i + 1,n_i) + Q_{m_i,n_i+1} w_i(m_i,n_i + 1) + Q_{m_i+1,n_i+1} w_i(m_i + 1,n_i + 1) \,.$$

This is an approximation of (9) and can be plugged directly into (6). The complexity of the quantization is linear in the number of samples, thus the complexity of the estimation is $O(N + B^2 \log B^2)$. Unlike estimates of discrete entropy, the estimate of empirical entropy is not dependent on the choice of B^2, since the summation over probabilities is carried out over the sample positions, rather than the bins. The choice does, however, influence the accuracy of the approximation.

Shwartz et al. (2005) also provides a gradient of the marginal entropy estimate, which we have generalized to joint entropy. The marginal entropy gradient is given with respect to the samples $\frac{\partial H}{\partial(\mathbf{a}^T\mathbf{X})}$. For the purpose of canonical information analysis the gradient with respect to the linear weighting $\mathbf{a}$ is needed. The chain rule yields

$$\frac{\partial h_x}{\partial \mathbf{a}} = \frac{\partial h_x}{\partial(\mathbf{a}^T\mathbf{X})}\frac{\partial(\mathbf{a}^T\mathbf{X})}{\partial \mathbf{a}} = \frac{\partial h_x}{\partial(\mathbf{a}^T\mathbf{X})}\mathbf{X}^T \ .$$

This is completely analogous for joint entropy estimation and the reader is referred to Shwartz et al. (2005) for further details.

The computational complexity of the approximate gradient estimation is of the order $O(B^2 \log B^2 + NN_{\text{dim}})$ where N_{dim} is the dimensionality of the linear weighting, i.e., either k or ℓ. In comparison, explicit calculation of the entropy gradient is of complexity $O(N_{\text{dim}}N^2 + N)$ (Shwartz et al., 2005).

6. Maximization of Mutual Information

The kernel density estimates of one- and two-dimensional pdfs by means of the method sketched in Section 4 are independent of additive and multiplicative transformations of each of the original variables. Therefore the maximization of the mutual information between the two linear combinations can be carried out without constraints. This means that very many optimization schemes may be applied.

Maximization of mutual information is inherently non-convex. For problems where it is not crucial to converge to the global optimum we suggest to use a local solver, e.g., either the downhill simplex method (Nelder and Mead, 1965) or Newton's method with the BFGS update (Fletcher, 1970), depending on whether one wishes to rely purely on function values or leverage the gradient introduced above. For problems where convergence to the global optimum is important, we propose to use a genetic algorithm at the cost of significantly more function evaluations. Results shown below are obtained using the genetic algorithm implemented in MATLAB with a population size of $5(k + \ell)^2$.

The choice of starting point is crucial when using local methods for global optimization. We have experimented with two different sets of starting points for each case, one being the optimum determined by canonical correlation analysis. The second set of starting points is constructed by letting $\mathbf{a}_0$ and $\mathbf{b}_0$ be unit vectors of length k and ℓ respectively, with an equal weighting on all variables, such that

$$\mathbf{a}_0 = \frac{1}{\sqrt{k}}\mathbf{1}_k, \ \mathbf{b}_0 = \frac{1}{\sqrt{\ell}}\mathbf{1}_\ell \tag{11}$$

where $\mathbf{1}_n$ is an n-vector of ones. For some problems, several candidate starting points may exist in which case we suggest to employ an optimization strategy where multiple local solvers start from individual starting points.

7. Case Studies

Here we give an illustrative toy example, an example which fuses weather radar and optical geostationary satellite data for a situation with heavy precipitation, and an example of using canonical information analysis for change detection in optical airborne data. These examples will be referred to as *toy*, *weather* and *cars* respectively for brevity.

The results are summarized in Table 2. Higher order components for these data sets were found to be trivial, wherefore only the leading component is shown.

7.1. Toy Example

In a simple, illustrative example consider the functions $f(x) = x$ and $g(x) = x^2$. The correlation between the functions over the interval [0,1] is $\sqrt{15/16} = 0.9682$, close to one. The correlation between the two over the interval [−1,1] is zero and yet of course the two variables are still closely associated.

Consider now this numeric example with a variable x_1 sampled equidistantly on the interval [0,1]. Let another variable x_2 be random Gaussian noise with mean zero and standard deviation one. Let y_1 be x_1^2 with random Gaussian noise with mean zero and standard deviation one tenth added. Let y_2 be random Gaussian noise with mean zero and standard deviation one. For all variables we have 1000 samples. Let the first set of variables consist of x_1 and x_2, and the second set consist of y_1 and y_2. In this case the leading canonical correlation is 0.9166 and (after sphering the input) the leading eigenvector for the first set is [1.0000 0.0064] and for the second set [1.0000 0.0143]. So in this case canonical correlation analysis makes sense: we get a high canonical correlation and eigenvectors that isolate the signal in x_1 and y_1. Maximal mutual information is 0.7867 and the leading projection vectors are [1.0000 0.0075] and [1.0000 −0.0043] respectively.

Let us now redo the analysis with x_1 sampled equidistantly on the interval [−1,1]. In this case the leading canonical correlation is 0.0532 and the leading eigenvector for the first set is [0.0391 0.9992] and for the second set [−0.8955 0.4450]. In this case canonical correlation analysis makes no sense: we get a very low canonical correlation and eigenvectors that do not isolate the signal in x_1 and y_1. Here maximal mutual information is 0.5856 and the leading projection vectors are [1.0000 −0.0082] and [1.0000 −0.0086] respectively.

For the latter case (x_1 sampled equidistantly on the interval [−1,1]), three-dimensional contour and scatter plots of the leading canonical variates are shown in Figures 2a (correlation based) and 2b (mutual information based). Figure 2a reveals no structure but in Figure 2b we clearly recognize the noisy parabola originally in variables x_1 and y_1. Unlike maximization of correlation of linear combinations of the two sets of variables, maximization of mutual information gives meaningful results in both cases.

We compare CIA to the 'explicit' (e.g., Yin (2004)) estimation of maximal mutual information projections performance in terms of accuracy and computation time. The accuracy is evaluated in terms of the geometric mean

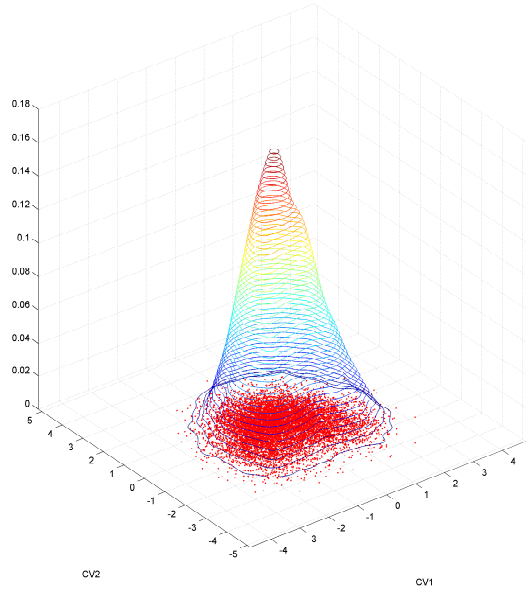
$$\mu = \sqrt{|\rho_1||\rho_2|} \tag{12}$$

of the absolute correlations $\rho_1 = \mathrm{corr}(x_1, U)$ and $\rho_2 = \mathrm{corr}(y_1, V)$. E.g., the correct $\mathbf{a}^\star = \mathbf{b}^\star = [1,0]^T$ would yield $\rho_1 = \rho_2 = \mu = 1$. Figure 3a shows the difference in geometric mean $\mu_{\mathrm{CIA}} - \mu_{explicit}$ for three different sample sizes $N = \{500,1000,5000\}$ and for ten values of the standard deviation σ for the noise added to x_1^2 to form y_1. We see that in low-noise cases ($\sigma < 0.6$) the difference in geometric mean is negligible, while both estimation procedures have difficulties for larger noise levels and sample sizes < 5000. Figure 3b shows the computation times as a ratio ('explicit'/CIA) of the time it has taken the genetic algorithm to converge. Note that the y-axis is in logarithmic units. For a sample size of $N = 500$ the speed is comparable, slightly in favor of the explicit estimation, for $N = 1000$ CIA is 1.4 times faster and for $N = 5000$ it is approximately 20 times faster. To put the computation time ratio into perspective, we note that for, e.g., $\sigma = 0.89$ and $N = 5000$ the explicit estimation takes 194.5 minutes to converge, while CIA takes 3.8 minutes to converge to an equally good solution with an average of 18.37 seconds and 0.36 seconds per function evaluation respectively. In the supplementary material we supply similar comparison plots for three other simulation scenarios suggested by Yin (2004).

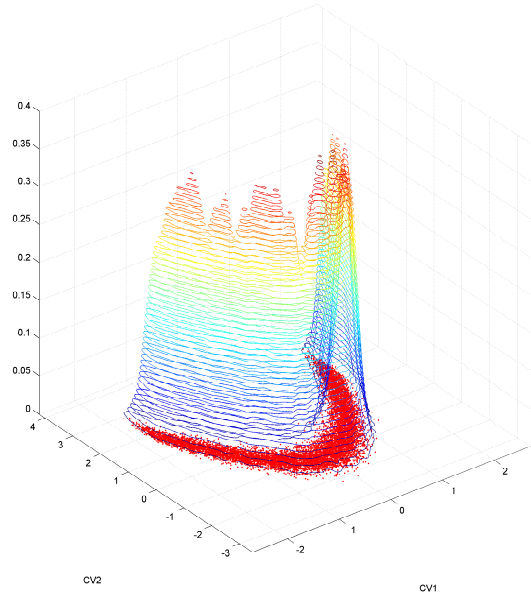
7.2. Weather Radar and Meteosat Data

This data set consists of satellite and radar imagery from 20 August 2007, where extreme downpour intensities (53 millimeter in 10 minutes) were recorded in some regions of Denmark.

The satellite imagery is a set of $k = 8$ infrared bands from the Spinning Enhanced Visible and Infrared Imager (SEVIRI) onboard the Meteosat Second Generation (MSG-2) weather satellite. The spectral region of the infrared bands are from approximately 3.9μm to 13.4μm, and these bands monitor cloud top reflectance properties. The radar data are recorded three minutes before the satellite image using the Danish Meteorological Institute (DMI) weather radars and consists of a single ($\ell = 1$) image of radar reflectance. The two image sources are gridded as images of 400 × 500 pixels with a ground sampling distance of 2 km × 2 km prior to analysis to establish pixel-to-pixel correspondence. The analysis includes the $N = 7{,}577$ observations in the radar imagery exhibiting reflectance from precipitation.



(a) Canonical correlation analysis



(b) Canonical information analysis

Figure 2: **Toy example** a) Correlation based canonical variates and b) mutual information based canonical variates for toy example with variables sampled equidistantly on the interval [−1,1].

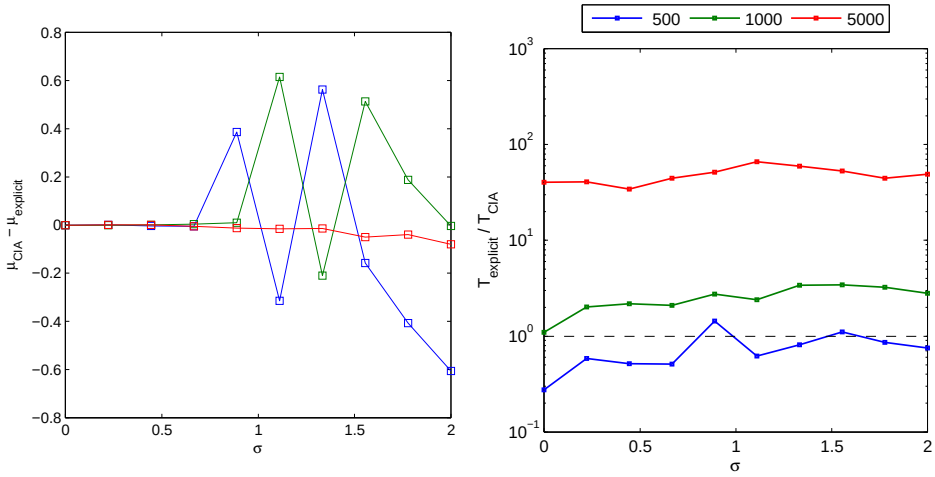


Figure 3: **Simulation studies** Comparison of accuracy and speed for CIA and 'explicit' estimation. σ is the noise level, μ_{CIA} and $\mu_{explicit}$ are defined as in Eq. (12) Values above 0 indicate a higher correlation between the found components and the true components (a better solution) for CIA, while values below 0 indicate a better solution yielded by the explicit estimation. Speed is shown as $\frac{T_{\text{explicit}}}{T_{\text{CIA}}}$ on a logarithmic scale, thus CIA is slower for values below 1 and faster for values above 1. The three colored lines represent results obtained with N = {500,1000,5000} simulated observations.

This case has also been treated by Vestergaard and Nielsen (2012), where an elaborate geometric and temporal alignment was needed to ameliorate the CCA solution. As will be shown below, this is entirely unnecessary when using the method suggested here.

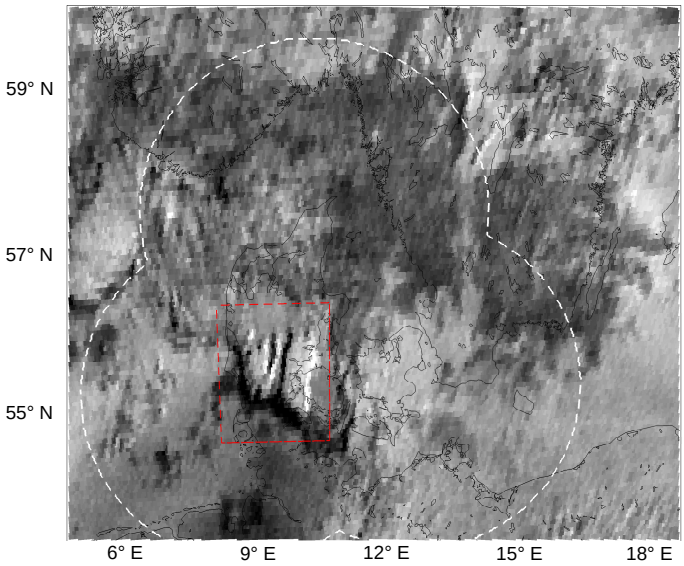
The motivation for fusion of these two data sources is twofold: First, weather radars have a limited coverage of approximately 240 kilometers from their position while satellites cover almost the entire planet. A fusion of these two could be a way of using satellite data as a proxy for radar data. Second, the two types of data come from very different types of sensors, wherefore the distributions of the data are very different. Therefore this is an illustrative example of using an information theoretic approach rather than a method based on assumptions of distributions.

The first mutual information canonical variate (MICV) is shown in Figure 4b where the eight infrared bands from the satellite data are projected onto the projection direction **a** determined by canonical information analysis. As the second set of variables consists of only a single variable, $\mathbf{b} = b = 1$. Therefore only the MICV related to the satellite data is shown. For comparison, the solution to the same problem determined by canonical correlation analysis is shown in Figure 4a. An area has been marked with a dashed red rectangle in both figures; extreme precipitation is known to occur in the dark region inside the rectangle in Figure 4b. A viable solution would therefore accentuate the cloud tops in this particular area. It is seen that this is the case for canonical information analysis, where a contrast with the surroundings is evident, while the correlation based result shows less contrast.

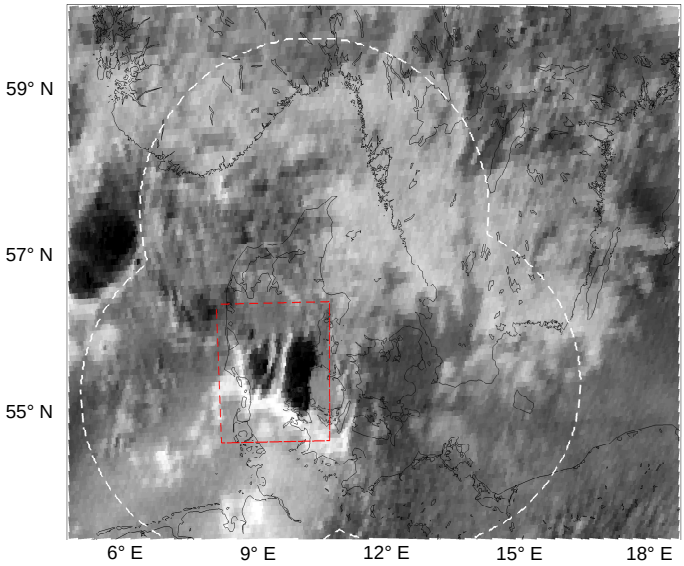
A correlation of 0.344 and 0.303 between the leading pair of canonical variates was obtained using CCA and CIA respectively. Mutual information between the two mutual information based canonical variates is 0.101 while it is 0.088 between the two correlation based variates.

Quantitative comparison of correlation based and mutual information based analysis can, for example, be done by calculating spatial autocorrelation over the marked region in Figures 4a and 4b. We have chosen to calculate the autocorrelation over spatial lags of $[0\ 1], [1\ 1], [1\ 0]$ and $[-1\ 1]$ to capture spatial correspondences in all directions. For both analysis methods these values are shown in Table 1.

The average value for the mutual information based analysis is 0.950 compared to 0.897 for the correlation based analysis. These values confirm the subjective evaluation that the spatial coherence is larger in the mutual information



(a) Canonical correlation analysis



(b) Canonical information analysis

Figure 4: **Weather** The first CV determined by canonical correlation analysis and canonical information analysis for the *weather* data set. The marked rectangular area is known – from radar imagery – to exhibit extreme rain at this particular point in time. The display range of the intensity values is within ± three standard deviations of the mean. The dashed white line marks the extent of the radar coverage.

Table 1: Results for the *weather* data set evaluated in terms of spatial autocorrelation in the region of interest.

Method	→	↘	↓	↙	Average
CIA	0.973	0.932	0.950	0.943	**0.950**
CCA	0.952	0.859	0.892	0.886	**0.897**

based solution compared to the correlation based analysis.

7.3. DLR 3K Data

The images used in this example were recorded with the airborne DLR 3K camera system (Kurz et al., 2007a,b) from the German Aerospace Center, DLR. This system consists of three commercially available 16 megapixel cameras arranged on a mount and a navigation unit with which it is possible to record time series of images covering large areas at frequencies up to 3 Hz. The 1000 rows by 1000 columns example images acquired 0.7s apart cover a busy motorway. These data have previously been treated by Nielsen and Canty (2009); Nielsen (2011) and Nielsen (2007) where the original RGB images can be seen. The data at the two time points were orthoprojected using global positioning system/inertial measurement unit (GPS/IMU) measurements and a digital elevation model (DEM). For flat terrain like here one pixel accuracy was obtained. In these data, the change occurring between the two time points will be dominated by the movement of the cars on the motorway. Undesired, apparent change will occur due to the movement of the aircraft and the different viewing positions at the two time points.

Figure 5b shows the difference image between the first set of MICVs whereby canonical information analysis acts as a tool for change detection. Previously, a method for change detection based on canonical correlation analysis has been proposed (Nielsen et al., 1998). Comparing with the solution obtained by canonical correlation analysis in Figure 5a it is evident that a much larger amount of change information is gained by using CIA: the background is much smoother and clearly distinguishable from the areas of change (the cars) and the extreme values are only present where change has actually occurred. The difference image between the second set of MICVs is included in the supplementary material. Since relevant changes are due to the moving cars on the motor way only, higher order CVs in this case do not contain further information.

To quantify the different quality of the solutions, a region in the difference image has been selected. This region is known not to change between the two acquisition times and is assumed to be constant over the region in an ideal difference image. The variance in this region will therefore represent the unwanted noise in the difference image and is denoted var(N) below. The ratio R between the signal-to-noise ratios for the two solutions is defined as

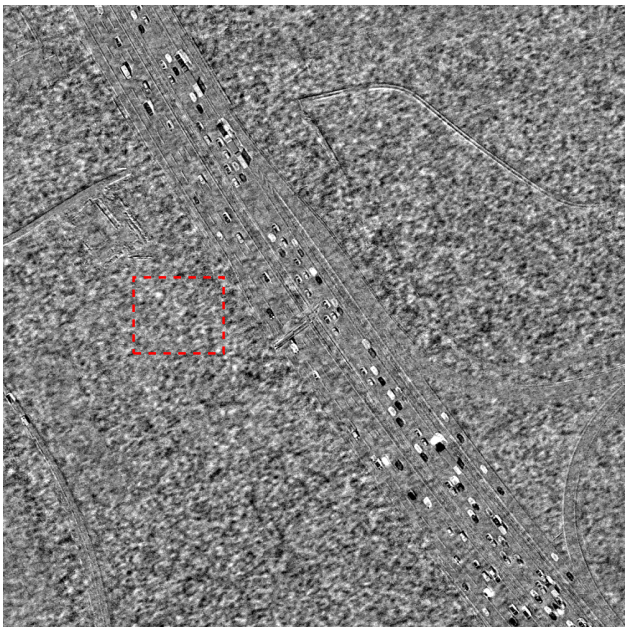
$$R = \frac{\mathrm{SNR}_{\mathrm{CIA}}}{\mathrm{SNR}_{\mathrm{CCA}}} = \frac{\frac{\mathrm{var}(S)}{\mathrm{var}(N_{\mathrm{CIA}})}}{\frac{\mathrm{var}(S)}{\mathrm{var}(N_{\mathrm{CCA}})}} = \frac{\mathrm{var}(N_{\mathrm{CCA}})}{\mathrm{var}(N_{\mathrm{CIA}})} \tag{13}$$

and is independent of the signal variance, when assuming that the true signal S is equal in the two solutions. The variance in this region for the solution produced by CIA is 0.265, while it is 0.878 for the correlation based solution, i.e., $R = 3.319$. This verifies the subjective evaluation that a more homogeneous no-change background is obtained using the proposed mutual information based method.

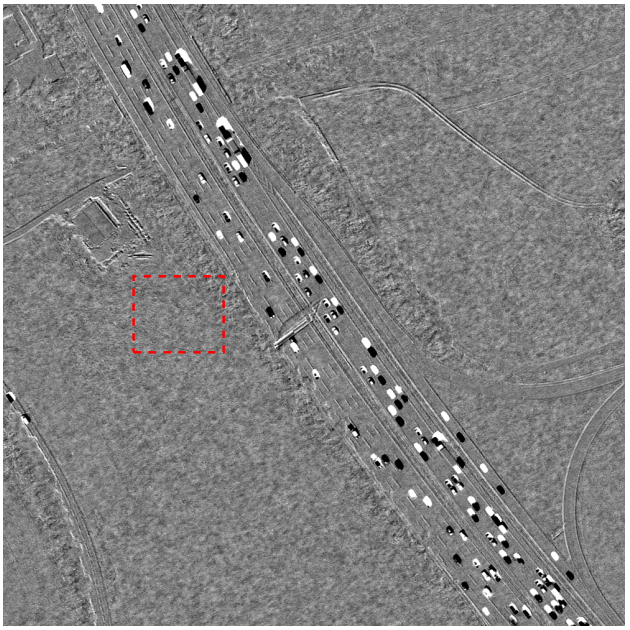
A correlation of 0.982 and 0.945 between the leading pair of canonical variates was obtained using CCA and CIA respectively, which demonstrates that a high correlation is not always the best measure for similarity. A mutual information of 1.034 and 1.335 between the leading pair of canonical variates was obtained using CCA and CIA respectively.

7.4. Summary

Table 2 summarizes the results for all three cases using canonical information analysis. Co-inspection of table and Figures 2, 4 and 5 clearly shows that the solution with the largest mutual information is superior to that with the largest correlation. Second order MICVs, MI between input bands and MICVs and a matrix of MI between pairs of MICVs are included as supplementary material for the *weather* and cars cases. Additional simulation studies suggested by Yin (2004) are detailed in the supplementary material, where the geometric mean using CIA, explicit estimation or CIA are shown.



(a) Canonical correlation analysis



(b) Canonical information analysis

Figure 5: **Cars** Difference image of the first set of MICVs for the *cars* data set using a) canonical correlation analysis and b) canonical information analysis respectively. The display range of the intensity values is within ± three standard deviations of the mean. The marked region is used to quantify the no-change noise variance.

In the *weather* case all 7,577 observations having a value in the radar data were used, while a random sample of 10,000 observations were used in the *cars* case were used for the optimization of mutual information. The determined linear transformations were applied to all observations in the two sets of variables. Each computation was done on a 64-bit Linux system with 2 X5650 6-Core processors, 2.66GHz, 48GB RAM.

Table 2: Summary of results for each of the three cases: *toy* is the toy example from Section 7.1, *weather* is the satellite/radar case from Section 7.2 and *cars* is the DLR 3K change-detection case from Section 7.3. I is mutual information as in Eq. (8), ρ is correlation, # is the number of function evaluations needed and sec. is the time in seconds..

		I	ρ	#	sec.
toyexample	CIA	0.127	0.010	4160	669
$(k,\ell) = (2,2)$	CCA	0.018	0.016	< 1	< 1
cars	CIA	1.335	0.945	9360	1165
$(k,\ell) = (3,3)$	CCA	1.034	0.982	< 1	< 1
weather	CIA	0.101	-0.303	21060	1672
$(k,\ell) = (8,1)$	CCA	0.088	0.344	< 1	< 1

In all three cases visual inspection of the resulting scatter plots and imagery clearly show the superior behavior of the mutual information based canonical analysis: the solution to the toy example illustrates that the CIA solution recovers the latent signal (the noisy parabola), while the CCA solution fails to do the same. The solution for the weather satellite data provides a representation of these data, which carry the most similar information to the weather radar data. This can be useful for, e.g., visualization purposes for meteorologists, or providing pseudo-radar coverage outside of the radar's range. In the change detection case, the background noise in the CCA solution looks almost similar to the signal, i.e., the cars. This is not the case for the CIA solution, where the noise in the difference image is suppresed and the cars stand out. This is clearly beneficial for any kind of application of these data.

8. Conclusions and Future Work

In this paper mutual information successfully replaces correlation to find canonical variates for two sets of multivariate observations. Unlike correlation which allows for second order statistics only, mutual information allows for the actual density of the variables at hand. An illustrative toy example with zero correlation between strongly associated variables proves the usefulness of the idea. Optical satellite data and weather radar data are successfully fused using the proposed method to accentuate precipitating clouds in the satellite data. This illustrates the benefit of mutual information when working with data sets of different modalities. Optical airborne (DLR 3K) data from two acquisition times 0.7s apart are included to illustrate the use of the proposed method in the context of change detection.

Canonical information analysis employs approximate marginal and joint entropy estimation. A simulation study shows that this approximation is as accurate as and much faster than previously presented algorithms, making the method feasible for image analysis problems and other large sample problems. Small sample applications ($N \leq 500$) do not benefit from this approach.

MATLAB software will be made available on the first author's homepage.

Acknowledgment

The authors would like to thank Researcher Dr. Thomas Bøvith and Aviation Meteorologist Birgitte Knudsen at Danish Meteorological Institute, DMI, for selecting and providing the weather radar and optical geostationary satellite data for the heavy precipitating case.

Thanks to Dr. Peter Reinartz and coworkers, German Aerospace Center, DLR, Oberpfaffenhofen, Germany, for letting us use the geometrically coregistered DLR 3K camera data

AAN initially started work on this subject during a sabbatical leave to the University of Oxford, Department of Statistics. Thanks to Professor Brian D. Ripley for hosting.

References

Bach, F. R., Jordan, M. I., 2002. Kernel independent component analysis. Journal of Machine Learning Research 3, 1–48.

Beirlant, J., Dudewicz, E. J., Györfi, L., Van der Meulen, E. C., 1997. Nonparametric entropy estimation: An overview. International Journal of Mathematical and Statistical Sciences 6, 17–40.

Bishop, C. M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st Edition. Springer.

Botev, Z. I., Grotowski, J. F., Kroese, D. P., Oct. 2010. Kernel density estimation via diffusion. The Annals of Statistics 38 (5), 2916–2957.

Canty, M. J., 2010. Image Analysis, Classification, and Change Detection in Remote Sensing, 2nd Edition. CRC Press/Taylor and Francis.

Conese, C., Maselli, F., 1993. Selection of optimum bands from {TM} scenes through mutual information analysis. {ISPRS} Journal of Photogrammetry and Remote Sensing 48 (3), 2 – 11.

de Bie, T., de Moor, B., 2002. On two classes of alternatives to canonical correlation analysis, using mutual information and oblique projections. In: Proceedings of the 23rd symposium on information theory in the Benelux (ITB). Louvain-la-Neuve, Belgium.

Ehlers, M., 1991. Multisensor image fusion techniques in remote sensing. {ISPRS} Journal of Photogrammetry and Remote Sensing 46 (1), 19 – 30.

Erten, E., Reigber, A., Ferro-Famil, L., Hellwich, O., 2012. A new coherent similarity measure for temporal multichannel scene characterization. IEEE Transactions on Geoscience and Remote Sensing 50 (7), 2839–2851.

Fletcher, R., 1970. A new approach to variable metric algorithms. The computer journal.

Friedman, J., 1987. Exploratory projection pursuit. Journal of the American Statistical Association 82 (397), 249–266.

Haber, E., Modersitzki, J., 2007. Intensity gradient based registration and fusion of multi-modal images. Methods of information in medicine 46 (3), 292–299.

Hotelling, H., Dec. 1936. Relations between two sets of variates. Biometrika 28 (3-4), 321–377.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent component analysis. Adaptive and learning systems for signal processing, communications, and control. J. Wiley.

Jones, M., Marron, J., 1996. A brief survey of bandwidth selection for density estimation. Journal of the American Statistical 91 (433), 401–407.

Karasuyama, M., Sugiyama, M., 2012. Canonical dependency analysis based on squared-loss mutual information. Neural networks 34, 46–55.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Physical Review E 066138 (May 2003), 1–16.

Kurz, F., Charmette, B., Suri, S., Rosenbaum, D., Spangler, M., Leonhardt, A., Bachleitner, M., Stätter, R., Reinartz, P., 2007a. Automatic traffic monitoring with an airborne wide-angle digital camera system for estimation of travel times. Science 36, 09–19.

Kurz, F., Müller, R., Stephani, M., Reinartz, P., Schroeder, M., 2007b. Calibration of a wide-angle digital camera system for near real time scenarios. In: Proc of ISPRS Hannover Workshop 2007-High Resolution Earth Imaging for Geospatial Information. pp. 1682–1777.

Lai, P. L., Fyfe, C., 2000. Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems 10 (5), 365–377.

Mackay, D. J. C., Jun. 2003. Information Theory, Inference and Learning Algorithms, 1st Edition. Cambridge University Press.

Modersitzki, J., 2004. Numerical methods for image registration. Oxford University Press, USA.

Nelder, J. A., Mead, R., 1965. A Simplex Method for Function Minimization. The Computer Journal 7 (4), 308–313.

Nielsen, A. A., Feb. 2007. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. Image Processing, IEEE Transactions on 16 (2), 463–78.
URL `http://www.imm.dtu.dk/pubdb/p.php?4695`

Nielsen, A. A., 2011. Kernel Maximum Autocorrelation Factor and Minimum Noise Fraction Transformations. Image Processing, IEEE Transactions on 20 (3), 612–624.
URL `http://www.imm.dtu.dk/pubdb/p.php?5925`

Nielsen, A. A., Canty, M. J., 2009. Kernel principal component and maximum autocorrelation factor analyses for change detection. Proceedings of SPIE 7477, 74770T–74770T–6.
URL `http://www.imm.dtu.dk/pubdb/p.php?5757`

Nielsen, A. A., Conradsen, K., Simpson, J., 1998. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. Remote Sensing of Environment 64 (1), 1–19.
URL `http://www.imm.dtu.dk/pubdb/p.php?1220`

Parzen, E., 1962. On estimation of a probability density function and mode. The annals of mathematical statistics 33 (3), 1065–1076.

Pohl, C., Van Genderen, J. L., 1998. Review article multisensor image fusion in remote sensing: Concepts, methods and applications. International Journal of Remote Sensing 19 (5), 823–854.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 832–837.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal 27 (3), 379–423 and 623–656.

Sheather, S., Jones, M., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), 683–690.

Shwartz, S., Zibulevsky, M., Schechner, Y., May 2005. Fast kernel entropy estimation and optimization. Signal Processing 85 (5), 1045–1058.

Silverman, B. W., 1986. Density estimation for statistics and data analysis. Vol. 26. Chapman & Hall/CRC.

Studholme, C., Hill, D., Hawkes, D., Jan. 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition 32 (1), 71–86.

Suri, S., Reinartz, P., Feb. 2010. Mutual-Information-Based Registration of TerraSAR-X and Ikonos Imagery in Urban Areas. IEEE Transactions on Geoscience and Remote Sensing 48 (2), 939–949.

Terrell, G., 1990. The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85 (410), 470–477.

Vestergaard, J. S., Nielsen, A. A., 2012. Automated invariant alignment to improve canonical variates in image fusion of satellite and weather radar data. Journal of Applied Meteorology and Climatology.
URL `http://journals.ametsoc.org/doi/abs/10.1175/JAMC-D-12-05.1`

Viola, P., 1997. Alignment by maximization of mutual information. International journal of computer vision 24 (2), 137–154.

Viola, P. A., 1995. Alignment by maximization of mutual information. Ph.D. thesis, Massachusetts Institute of Technology.

Yin, X., 2004. Canonical correlation analysis based on information theory. Journal of Multivariate Analysis 91, 161–176.

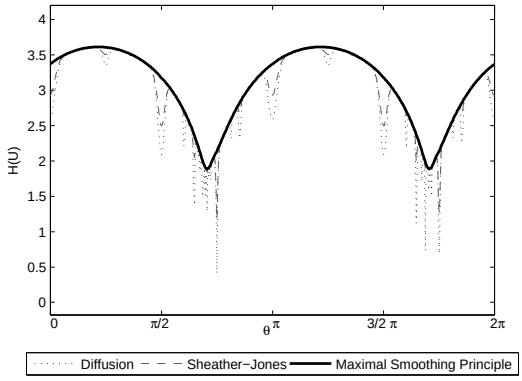


Figure A.6: Comparison of entropy estimates based on kernel density estimates using three different bandwidth estimators: A diffusion based estimator, the "solve-the-equation plug-in" estimator by Sheather-Jones and the maximal smoothing principle.

Appendix A. Comparison of bandwidth estimators

Here we motivate the choice of the maximal smoothing principle (Terrell, 1990) for kernel bandwidth estimation by comparing its properties with the Sheather-Jones estimator (Sheather and Jones, 1991) and a diffusion based estimator (Botev et al., 2010).

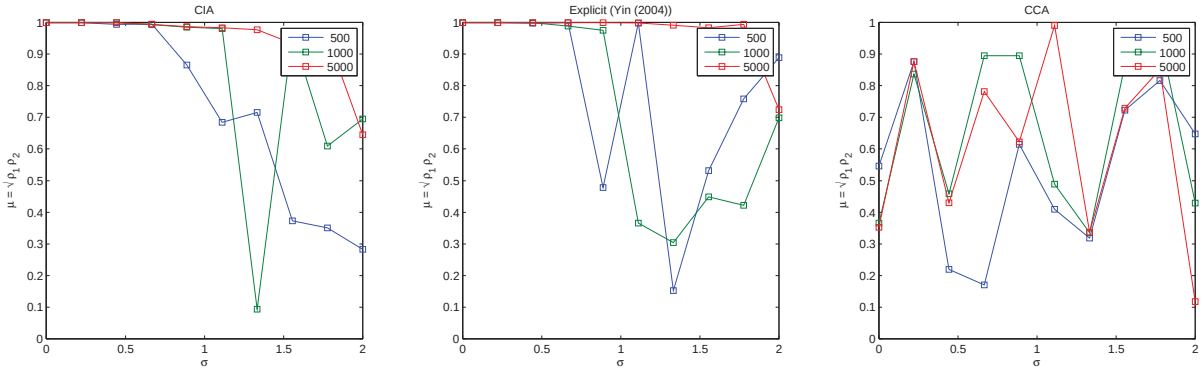
The desirable properties of the maximal smoothing principle for kernel bandwidth estimation can best be illustrated by an example. For illustration purposes we consider a single set of a two-dimensional stochastic variable $\mathbf{X}$. We wish to estimate the entropy of the linear combination $U = \mathbf{a}^T\mathbf{X}$ using a kernel density estimator. The entropy becomes a function of the bandwidth estimate $H(\hat{\sigma}_X(\mathbf{a}|\mathbf{X}))$. The bandwidth is estimated based only on the linear combination U and is thereby a function of the projection direction $\mathbf{a}$ given the data $\mathbf{X}$.

We let $\mathbf{a}$ be a vector on the unit circle and it can thus be fully described in spherical coordinates as $\mathbf{a}(\theta) = (1, \theta)$ by the angle θ. In the following experiment we vary the angle over the range $\theta \in [0, 2\pi]$ and estimate the bandwidth $\hat{\sigma}_X$ for each value of θ. This bandwidth is used for calculating the entropy.

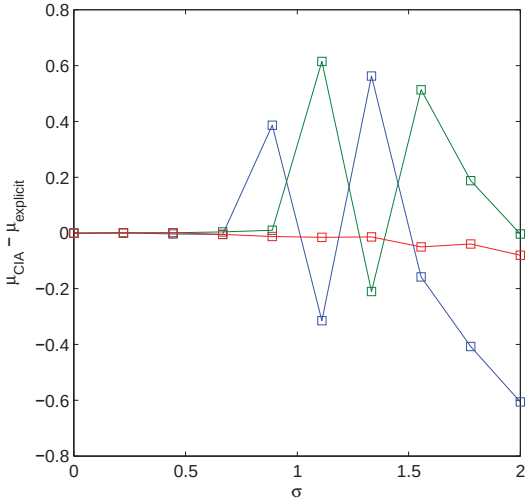
Figure A.6 shows the entropy $H(U)$ as a function of the projection direction angle θ for three different bandwidth estimators. It is immediately seen that the entropy estimate is smoother and avoids local minima when using the maximal smoothing principle, while the Sheather-Jones estimator and the diffusion based estimator fluctuate much more. The average computation times over 500 estimations of the bandwidth is 0.09, 72.03 and 0.04 seconds for the diffusion based based estimator, the Sheather-Jones and the maximial smoothing prinicpal, respectively.

Based on these observations, we find the maximal smoothing principle best suited for estimation of bandwidth in the context of optimizing mutual information of linear combinations. Though this behavior is illustrated in two-dimensional data only, we employ this principle for higher dimensional data as well.

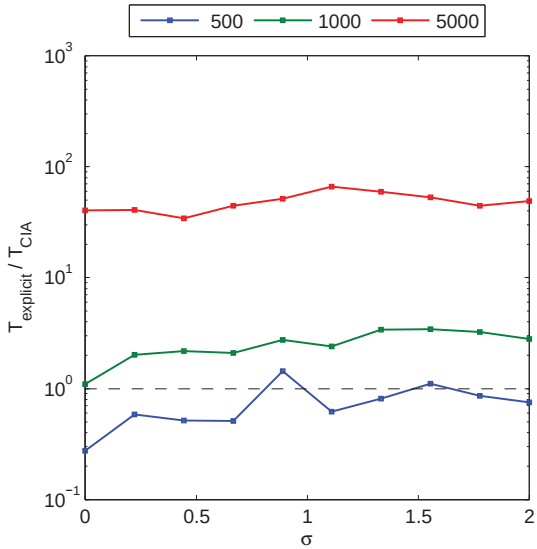
Example 1/Toy example



(a) Geometric mean for CIA, explicit estimation and CCA.



(b) Geometric mean differences.

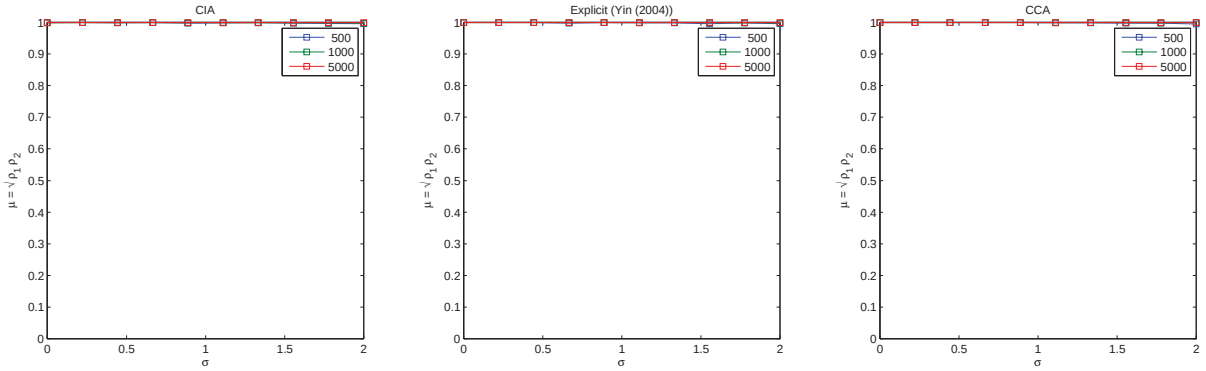


(c) Time comparison.

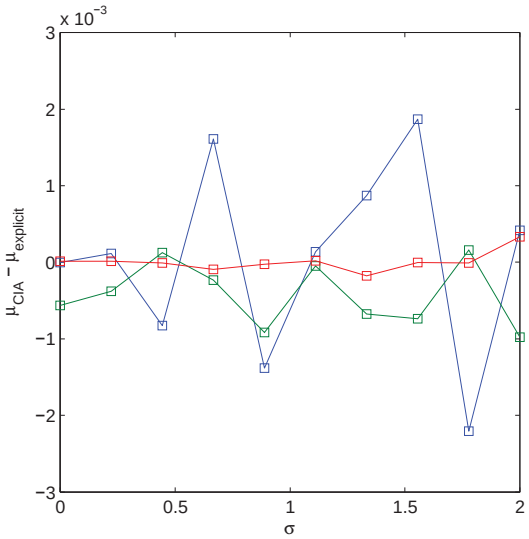
Distributions of simulated variables

$x_1 \sim U(-1,1)$	$y_1 = x_1^2 + \sigma\varepsilon$
$x_2 \sim N(0,0.1^2)$	$y_2 \sim N(0,0.1^2)$
	$\varepsilon \sim N(0,1)$

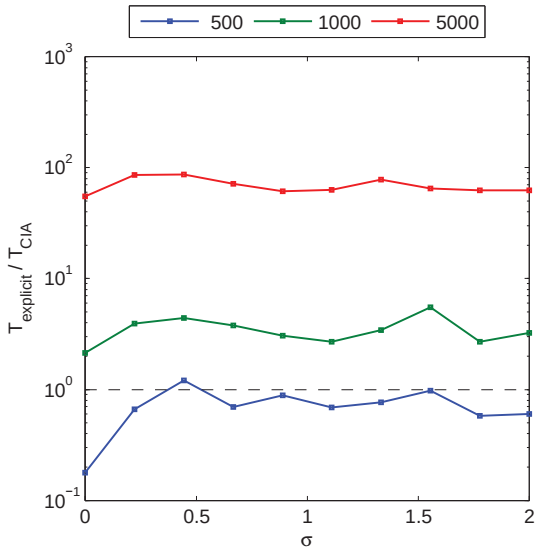
Example 2



(a) Geometric mean for CIA, explicit estimation and CCA.



(b) Geometric mean differences.

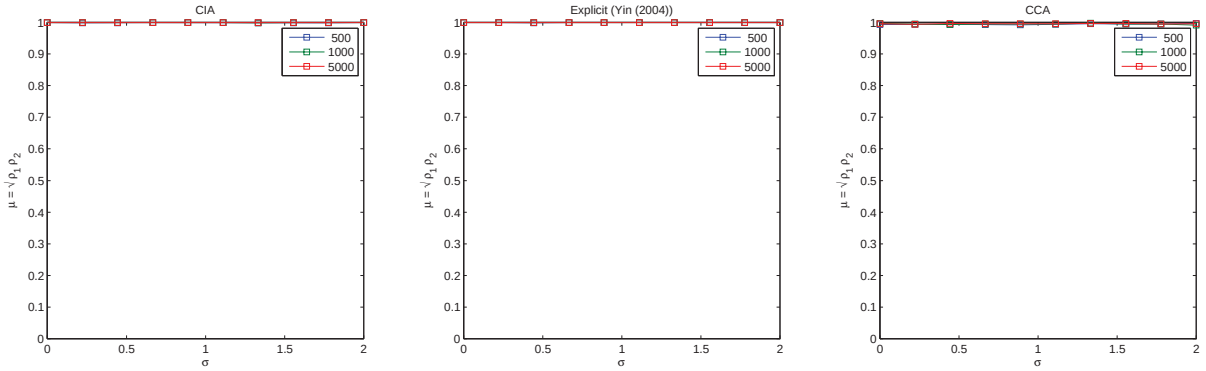


(c) Time comparison.

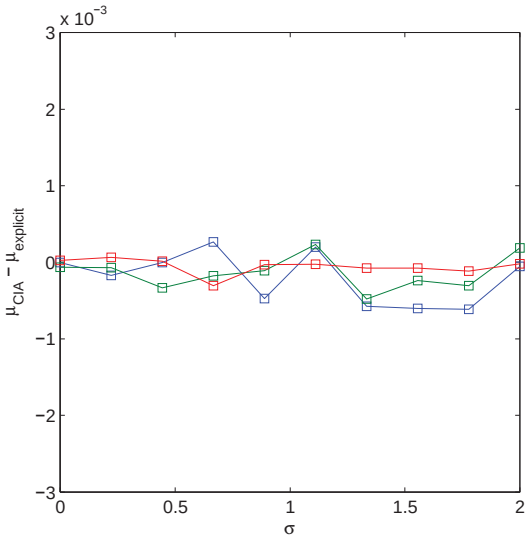
Distributions of simulated variables

$x_1 \sim U(0,1)$	$y_1 = x_2 + 2x_3 + 2x_4 + \sigma\varepsilon_1$
$x_2 = 1 + x_1^2 + 0.02\varepsilon$	$y_2 \sim t(5)$
$x_3 \sim N(0,1)$	$y_3 \sim N(0,1)$
$x_4 \sim \chi^2(3)$	$\varepsilon,\varepsilon_1 \sim N(0,1)$
$x_5 \sim Gam(1,4)$	
$x_6 \sim N(0,1)$	

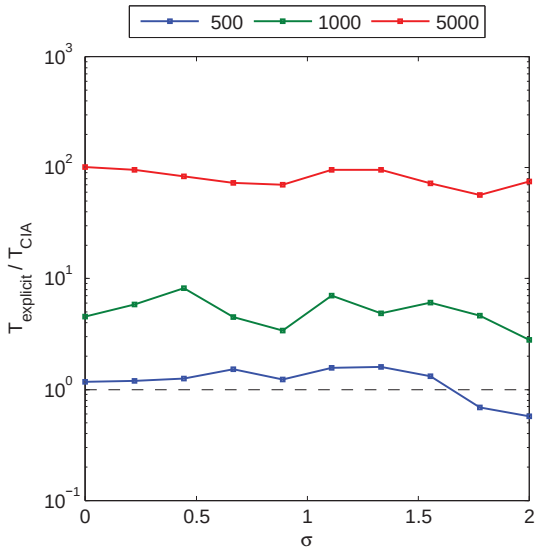
Example 3



(a) Geometric mean for CIA, explicit estimation and CCA.



(b) Geometric mean differences.

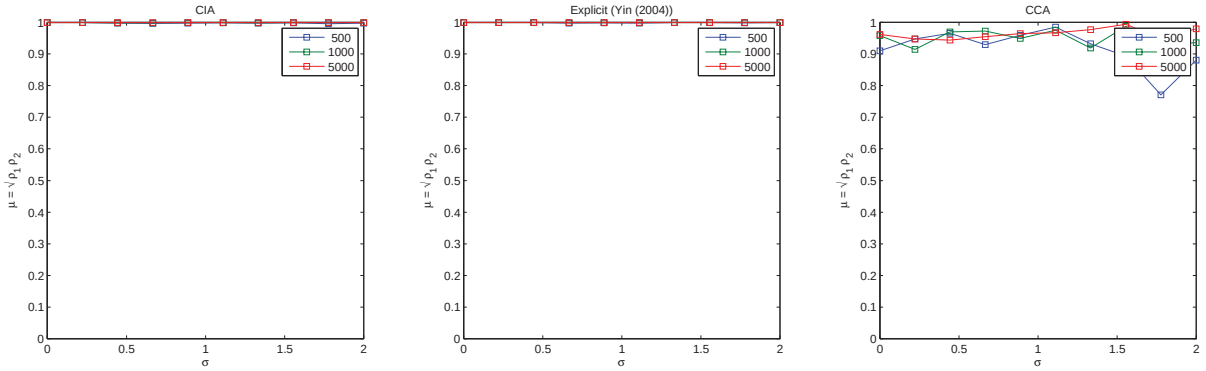


(c) Time comparison.

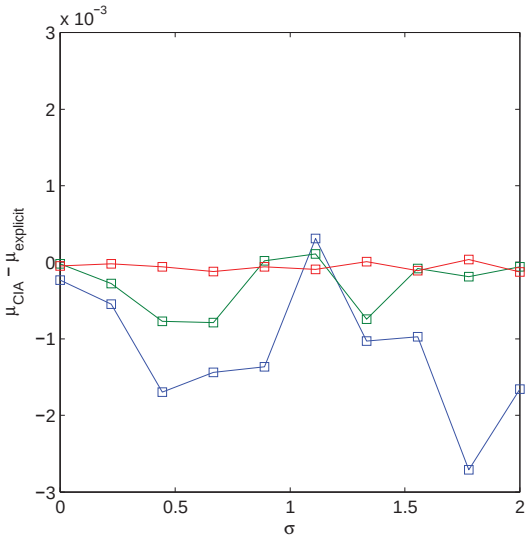
Distributions of simulated variables

$x_1 \sim t(6)$	$y_1 = (x_2 + 3x_3 + 2x_4)^2 + \sigma\varepsilon$
$x_2 \sim \chi^2(7)$	$y_2 \sim t(13)$
$x_3 \sim N(0,1)$	$y_3 \sim \chi^2(13)$
$x_4 \sim t(8)$	$\varepsilon \sim N(0,1)$
$x_5 \sim F(3,12)$	
$x_6 \sim \chi^2(3)$	
$x_7 \sim Gam(1,4)$	
$x_8 \sim N(0,1)$	
$x_9 \sim t(5)$	
$x_{10} \sim U(0,1)$	

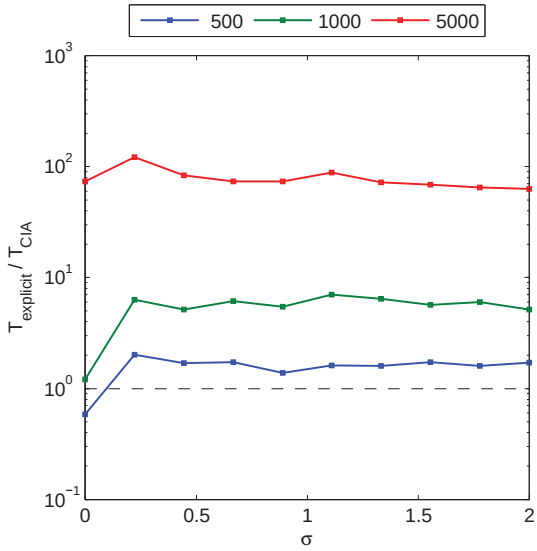
Example 4



(a) Geometric mean for CIA, explicit estimation and CCA.



(b) Geometric mean differences.



(c) Time comparison.

Distributions of simulated variables

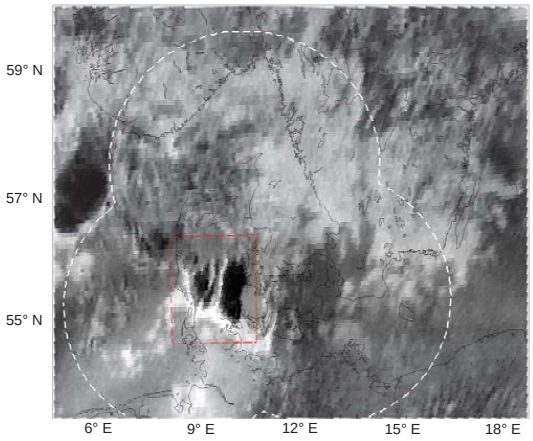
$x_1 \sim \chi^2(7)$	$y_1 = \sigma(x_2 + 3x_3 + 2x_4)^2\varepsilon$
$x_2 \sim N(0,1)$	$y_2 \sim t(6)$
$x_3 \sim t(8)$	$\varepsilon \sim N(0,1)$
$x_4 \sim F(3,12)$	
$x_5 \sim \chi^2(3)$	
$x_6 \sim Gam(1,4)$	
$x_7 \sim N(0,1)$	
$x_8 \sim t(5)$	

Weather

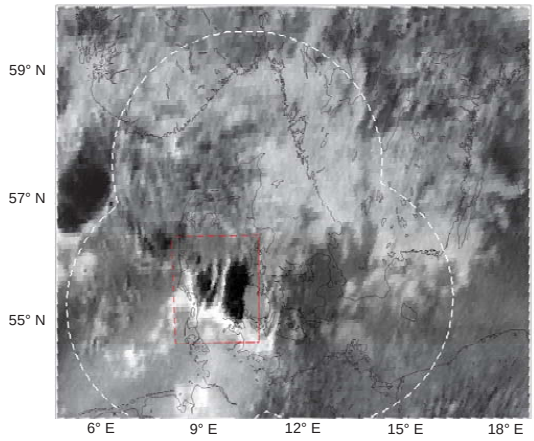
Two MICVs are shown for the *weather* case. Note that the second MICV is merely a noisy version of the first MICV. This is due to the main information at this point in time of the weather data is the extreme rain captured by the first component. The bar plots show MI between the input variables and the found components, e.g., $MI(x_i,U_1)$ in Figure (c) and $MI(y_i,V_1)$ in Figure (d). The very similar distribution of MI across the input variables for the two MICVs also suggests that the two MICVs are very similar in nature. Note that due to the non-negativity of MI, these association plots are unfortunately less informative than correlation plots usually used for CCA.

Table (b) shows the MI and correlation ρ for the first set of CVs. Table (c) shows the MI between all pairs of CVs. We see that $MI(U_1,V_1) > MI(U_2,V_2)$ meaning that the MICVs are automatically sorted properly due to the employed search strategy

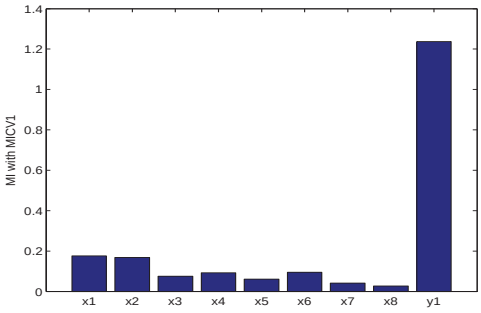
The radar data used as the second set of variables are in fact one dimensional. This has the implications that 1) CCA cannot determine a second component, since it requires orthogonality, and 2) $V_1 = V_2 = y_1$, which is also apparent in Table (c).



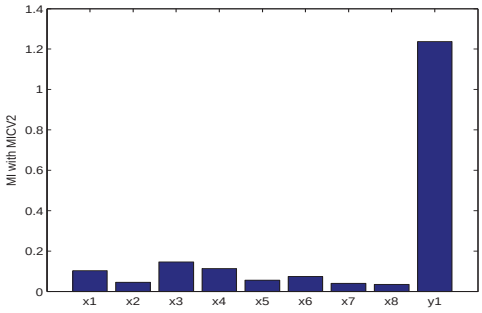
(a) MICV 1



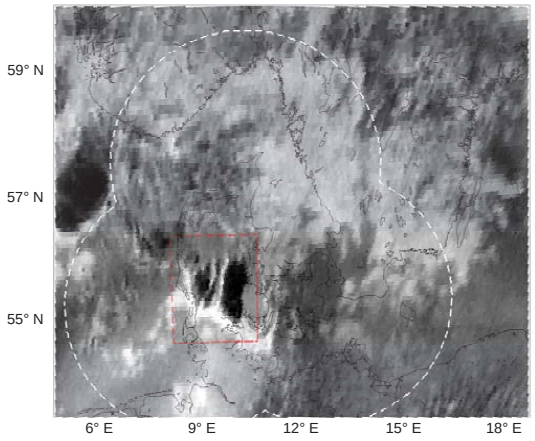
(b) MICV 2



(c) MI(data,MICV 1)



(d) MI(data,MICV 2)



(a) CCA CV 1

	MI	ρ
CCA	0.088	0.344
CIA	0.101	0.303

(b) MI and correlation for first CV.

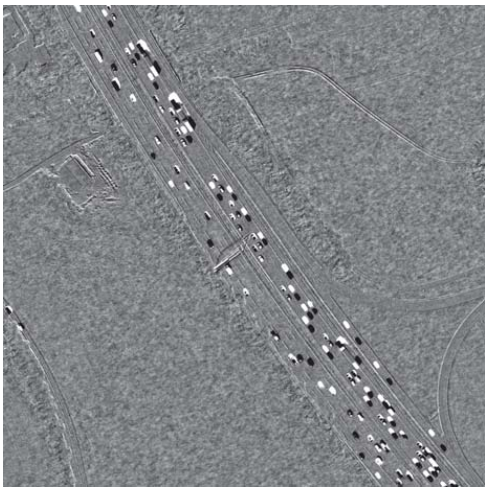
	U_1	U_2	V_1	V_2
U_1	1.879	0.182	0.101	0.101
U_2	0.182	1.786	0.028	0.028
V_1	0.101	0.028	1.479	1.479
V_2	0.101	0.028	1.479	1.479

(c) MI matrix

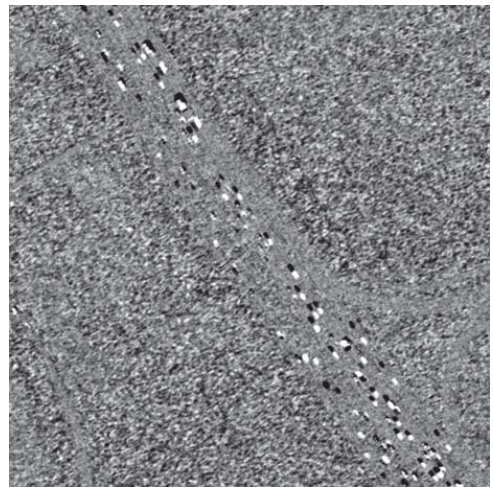
Cars

Two MICVs are shown for the *cars* case. Since relevant changes are due to the moving cars on the motor way only, higher order CVs in this case do not contain further information The bar plots show MI between the input variables and the found components, e.g., $MI(x_i,U_1$ and $MI(y_i,V_1)$ in Figure (c). It is seen that the MI between x_1,x_2,y_1,y_2 and the found components are strong in the first component and less so in the second component. Note that due to the non-negativity of MI, these association plots are unfortunately less informative than correlation plots usually used for CCA.

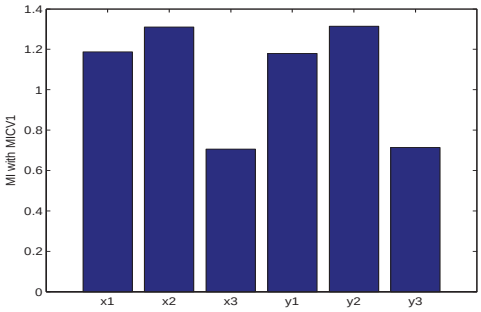
Table (b) shows the MI and correlation ρ for the first set of CVs. Table (c) shows the MI between all pairs of CVs. We see that $MI(U_1,V_1) > MI(U_2,V_2)$ meaning that the MICVs are automatically sorted properly due to the employed search strategy.



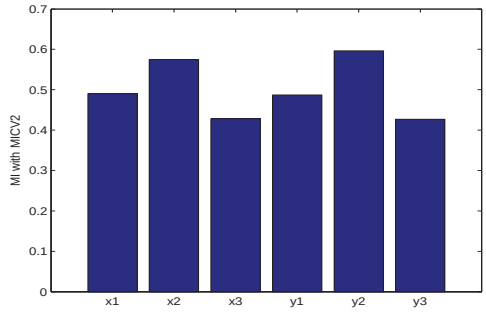
(a) Differences of MICV 1



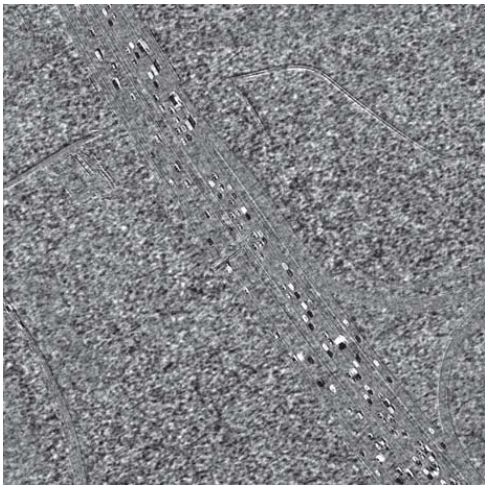
(b) Differences of MICV 2



(c) MI(data,MICV 1)



(d) MI(data,MICV 2)



(a) Differences of CCA CV 1

	MI	ρ
CCA	1.034	0.982
CIA	1.335	0.945

(b) MI and correlation for first CV.

	U_1	U_2	V_1	V_2
U_1	1.804	0.529	1.335	0.503
U_2	0.529	1.418	0.495	1.080
V_1	1.335	0.495	1.809	0.545
V_2	0.503	1.080	0.545	1.434

(c) MI matrix

Paper B

Number of genes controlling a quantitative trait in a hybrid zone of the aposematic frog Ranitomeya imitator

Number of genes controlling a quantitative trait in a hybrid zone of the aposematic frog *Ranitomeya imitator*

Jacob S. Vestergaard[a], Evan Twomey[b], Rasmus Larsen[a], Kyle Summers[b], Rasmus Nielsen[c]

[a]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Building 324/130, Matematiktorvet, 2800 Kgs Lyngby, Denmark. Phone: +4545253427.*
[b]*Department of Biology, East Carolina University, Howell Science Complex N314, Greenville, NC 27858*
[c]*Department of Integrative Biology, University of California Berkeley, 4098 VLSB, Berkeley, CA 94720*

Abstract

The number of genes controlling mimetic traits has been a topic of much research and discussion. In this paper we examine a mimetic, dendrobatid frog *Ranitomeya imitator*, which harbors extensive phenotypic variation with multiple mimetic morphs, not unlike the celebrated *Heliconius* system. However, the genetic basis for this polymorphism is unknown, and not easy to determine using standard experimental approaches, for this hard-to-breed species. To circumvent this problem, we first develop a new protocol for automatic quantification of complex color pattern phenotypes from images. Using this method, which has the potential to be applied in many other systems, we define a phenotype associated with differences in color pattern between different mimetic morphs. We then proceed to develop a maximum likelihood method for estimating the number of genes affecting a quantitative trait segregating in a hybrid zone. This method takes advantage of estimates of admixture proportions obtained using genetic data, such as microsatellite markers, and is applicable to any other system where a phenotype has been quantified in an admixture/introgression zone. We evaluate the method using extensive simulations and apply it to the *R. imitator* system. We show that likely one or two, or at most three genes, control the mimetic phenotype segregating in a *R. imitator* hybrid zone identified using image analyses.

Keywords: Ranitomeya imitator, hybridization, image analysis, quantitative phenotyping

1. Introduction

The analysis of phenotypic and genetic variation in geographic areas where two or more phenotypically distinguishable groups of organism meet and exchange genes has been of substantial interest to evolutionary biologists (see e.g., Endler, 1977; Barton and Hewitt, 1985; Coyne and Orr, 2004). The evolutionary dynamics in these zones, referred to as hybrid zones, introgression zones, or admixture zones depending on context, provide a basis for studying processes relating to speciation and for understanding the genetic and ecological underpinnings of adaptive traits, including mimetic and aposematic traits. Substantial work has been done on such systems, including the now classical work on the *Bombina bombina* vs. *B. variegata* hybrid zone (e.g., Szymura and Barton, 1991, 1986) and the hybrid zones between various species of *Heliconius* butterflies (e.g., Turner, 1971; Mallet, 1986; Jiggins et al., 2001). Of primary interest in these studies is to understand the genetic basis of the phenotypic traits, how selection is affecting these traits, and to understand the relative role of population history, gene-flow and natural selection in determining the evolutionary dynamics of the hybrid zone. Furthermore, there has recently been renewed interest in mapping the genetic variants associated with reproductive isolation or adaptive traits in the hybrid zone using so-called admixture mapping or mapping by admixture linkage disequilibrium (e.g., Chakraborty and Weiss, 1988; Briscoe et al., 1994; Patterson et al., 2004; Gompert and Buerkle, 2009; Winkler et al., 2010; Crawford and Nielsen, 2013). Of special interest to us is admixture zones, exemplified by the previously mentioned examples in *Bombina* and *Heliconius*, in which complex morphological traits such as color patterns are segregating and are likely of adaptive significance.

Mimetic traits in admixture zones, or otherwise, have often been hypothesized to be associated with so-called 'supergenes' (Clarke et al., 1968). Supergenes are tightly linked clusters of genes that control a suite of traits that will allow Mendelian, or close to, Mendelian behavior of the mimicry trait. The existence of such supergenes could help explain the strong phenotypic correlation between many different phenotypes required to produce pure mimetic forms. If many traits are needed to produce an adaptive mimetic phenotype, one would expect selection to favor genetic variants that increase the correlation between these traits. Much discussion has ensued on the existence of supergenes, particularly in relation to mimetic phenotypes in butterflies (e.g., Joron and Mallet, 1998).

A recent paper by Kunte et al. (2014) show that a proposed supergene underlying memetic phenotypes in *Papilio* butterflies in fact is a single Mendelian gene that serves as a genetic switch for the mimetic type. For both *Papilio* and *Heliconius* it appears that the mimetic phenotypes are often controlled by one or a few genes or supergenes that behave in a largely Mendelian fashion. However, the degree to which mimetic phenotypes have a similar genetic basis in other systems is uncertain.

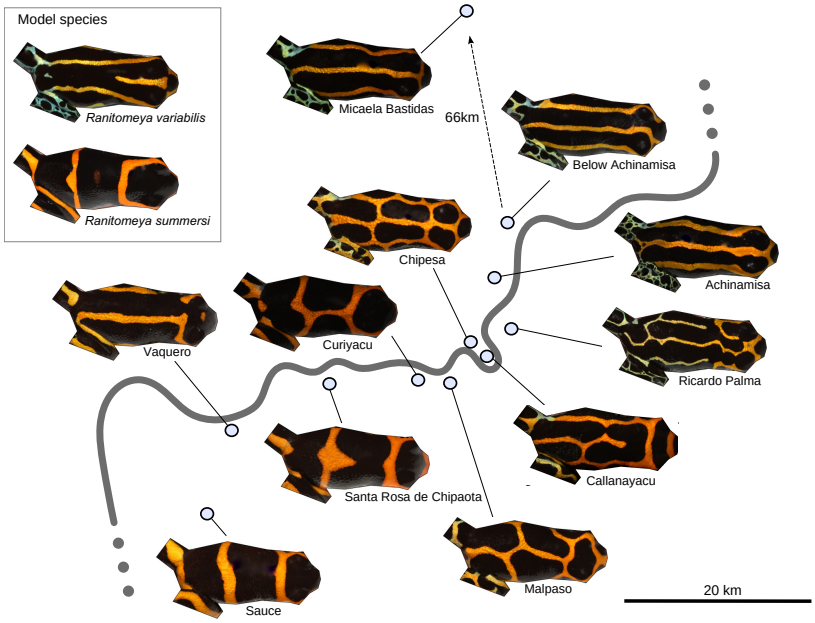


Figure 1: Sketch of sampling locations along the Huallaga River (grey). The two model species (*R. variabilis* and *R. summersi*) are shown in the upper left corner and examples of *R. imitator* are connected to their sampling localities.

The dendrobatid frog *Ranitomeya imitator* (Twomey et al., 2013a; Symula et al., 2001), provides a new vertebrate model system that shares many features with the well-known *Heliconius* system. Phylogenetic and phylogeographic analyses (Symula et al., 2001, 2003; Brown et al., 2011) indicate that this species is a member of a clade centered in southern Peru, but *R. imitator* is distributed in north-central Peru, in and around the province of San Martin. In this region, there are four distinct color pattern morphs of *R. imitator* that occupy different geographic regions (Yeager et al., 2012). In each of these regions, the color pattern of *R. imitator* clearly resembles that of a co-occurring species of dendrobatid frog (Symula et al., 2001; Twomey et al., 2013b). Phylogenetic analyses indicate that these co-occurring species generally diverged prior to the divergence between the divergent populations of *R. imitator* (Symula et al., 2003; Brown et al., 2011). Evidence for rapid divergence under selection (Symula et al., 2001; Yeager et al., 2012), and the similarity of each *R. imitator* color pattern morph to the more anciently diverged co-occurring species, indicates that *R. imitator* has undergone a mimetic radiation, in which different populations have evolved to resemble distinct color patterns displayed by the local model species (Symula et al., 2001; Yeager et al., 2012). In one case, *R. imitator* resembles two distinct color pattern morphs of a single species (*Ranitomeya variabilis*) that vary between highland (spotted) and lowland (striped) forms. In this case, it is not clear whether *R. imitator* adverged onto the color pattern of *R. variabilis*, or the reverse (Chouteau et al., 2011). However, in general the evidence supports the hypothesis that *R. imitator* adverged onto a co-occurring species, rather than the reverse (Symula et al., 2001, 2003; Yeager et al., 2012; Twomey et al., 2013b). Recent analyses of color pattern variation, genetic structure and gene flow have identified multiple zones of admixture where distinct color pattern morphs of *R. imitator* come into contact and interbreed (Twomey et al., 2013b). These regions vary in terms of the width of the zone of admixture and the degree of genetic divergence (in neutral markers) found across the zone, making this system useful for comparative analyses of divergence. In one region, the zone of admixture is fairly broad (7km), and populations in the zone show high variability that appears to include elements of both distinct color pattern morphs, see Figure 1. Hence interbreeding is likely to have proceeded for multiple generations in this zone of admixture, providing an excellent opportunity for infer-

ences concerning the genetic control of color pattern. Notice in Figure 1, that frogs at one end of the admixture zone, where they are mimetic with *R. summersi*, tend to be banded with black and orange legs, while frogs on the other end, where they are mimetic with *R. variabilis*, tend to be striped with a reticulated green and black pattern on the legs. The genetic basis of this polymorphism is of primary interest, but given the large genome sizes of dendrobatid frogs, lack of genetic resources, and difficulty of captive breeding, direct mapping of the genes involved is a non-trivial task. An objective of this paper is instead to obtain more information about the genetic basis of this polymorphism, solely using image analyses and limited microsatellite typing. In particular, we will be interested in examining if the polymorphism is controlled by a single Mendelian gene, perhaps a supergene, or by multiple genes. We will develop statistical methodologies for answering the question regarding the number of genes controlling the mimetic phenotypes and we will apply these methods to the *R. imitator* to determine the likelihood that the mimetic phenotypes in this system also are controlled by a supergene or a single Mendelian gene.

In order to address this problem, we will first develop automated methods for describing complex color pattern phenotypes based on images, that can be applied in this system and other systems. The advantage of such methods is that they are not subject to the same biases that may occur when a researcher chooses which traits to measure after having observed the images. In addition, such methods may have the potential for identifying important biological features that were otherwise not readily identifiable.

We will then proceed to develop a method for estimating the number of genes affecting a phenotype in an admixture/hybrid zone. For natural populations, in which controlled crosses are difficult or expensive to carry out, and for which parent-offspring pairs cannot easily be sampled, there are no appropriate methods for determining how many genes affect a trait. In other settings, there has been substantial previous work on this problem. The well-known Castle-Wright estimator (Castle, 1921; Wright, 1968) is based on the amount of segregating variation observed in the offspring of controlled crosses of inbred lines. The objective is to estimate the effective number of loci controlling a quantitative trait, i.e. the number of loci required to explain the variance in the trait if all loci have the same effect. There have been numerous extensions of the method including the incorporation of linkage and variation in effects size (e.g., Zeng, 1992; Otto and Jones, 2000). Lande (1981) showed that the assumption of complete homozygosity in the parental lines is not necessary and provided an estimator applicable to natural populations, rather than controlled crosses. Building on the idea, dating back to Pearson (1904), that the relationship between the variance in the offspring phenotypic values and midparent value depend on the number of genes controlling the trait, Slatkin (2013) provided another estimator applicable to outbred populations.

We are interested in estimating the number of genes affecting a trait in a hybrid/admixture zone. This is a problem that has been considered by Szymura and Barton (1991) who, based on theory developed in Barton (1983) and Barton and Bengtsson (1986), estimated the number of genes contributing to selection against gene-flow in the *Bombina bombina* vs. *B. variegata* hybrid zone using comparisons of the amount of linkage disequilibrium at the center of a hybrid zone to the width of the cline. The method we will develop is in the spirit of the of the Castle-Wright-Lande etimators, but is based on using a genetically inferred admixture proportion in each individual. This method does not require data on controlled crosses. It also does also not rely on any assumptions regarding selection models and processes shaping linkage disequilibrium. It is less ambitious in that it does attempt to determine the number of genes affecting fitness, but the number of genes affecting an observable phenotype. There is substantially less information regarding the number of loci when controlled crosses have not been performed. However, as we will show, there is still sufficient information to distinguish between hypotheses regarding a few, or many, genes affecting the trait.

We will apply these methods to images and genetic data from the aforementioned dendrobatid frog *Ranitomeya imitator*. Dendrobatid frogs typically have very large genome sizes (e.g. up to GB, Camper et al. (1993)), and genome sequencing and direct mapping using admixture or association mapping is difficult. Furthermore, experimental crosses and captive breeding can be challenging to carry out for these species. However, using the methods developed in this paper we can estimate the number of genes affecting the mimetic phenotype without the use of experimental crosses or mapping approaches.

2. Image analysis / quantitative phenotyping

A common way to quantify variation in image analysis is to extract a number of so-called descriptors, combine these into a vector of measurements for each individual and use statistical decomposition methods to condense the collected information. Prior to analysis all individuals have been warped to a mean shape determined by Procrustes analysis (Goodall, 1991). Manual annotation of 22 anatomical landmarks was used to establish point correspondences.

Descriptors are typically designed to capture elementary characteristics of an image, such as color or shape. Individually, descriptors are usually too specific, but a well-chosen suite of descriptors can provide a rich basis for further analysis.

In our study, we use three different phenotypic descriptors: color/non-color ratio, gradient orientation histograms and shape index histograms (Koenderink and van Doorn, 1992), each of which is defined on the pixel-level and described in detail in the electronic supplementary material ESM 1. These descriptors collect local 0th, 1st and 2nd order information about the image. In the current setting, these three standard descriptors can loosely be thought of as measuring features relating to the proportion of colored area, the degree to which changes in color occur along the anteroposterior axis or along the left-right axis (banded patterns versus striped patterns), and the degree to which the pattern consists of stripes/bands as opposed to reticulation, respectively. The quantified information is visualized in Figure 2 and in ESM 1.

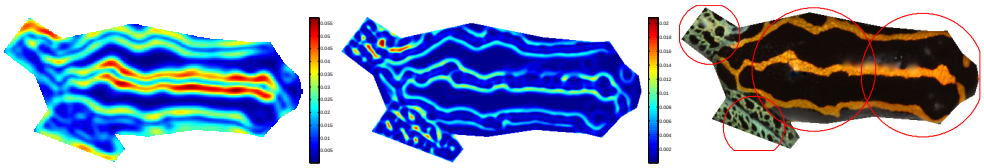
All descriptors are extracted on a per-pixel basis and pooled together at four distinct interest points, namely each of the frog's legs, lower back (dorsum) and on the back of head. An interest point is defined in terms of its coordinates $\mathbf{x}_k = [x_k, y_k]$ and a radius $r_k > 0$. Thus an average of each of the descriptors is accummulated for the four regions shown in Figure 2c.

This pooling scheme serves the purpose of reducing the number of descriptors extracted, without compromising the phenotypic variation captured. The pooling function is defined in terms of the k'th interest point as the circular average of a descriptor $g(\mathbf{x})$

$$P(\mathcal{Z}_k) = \frac{1}{|\mathcal{Z}_k|} \sum_{\mathbf{x}' \in \mathcal{Z}_k} g(\mathbf{x}') \tag{1}$$

where $\mathcal{Z}_k$ is the set of pixels with coordinates $\mathbf{x}'$ fulfilling $\|\mathbf{x}' - \mathbf{x}_k\| \leq r_k$.

The interest points and chosen radii are illustrated in Figure 2c.



(a) Gradient at $b = \pm\frac{\pi}{2}$.

(b) Shape index at $b = -0.8$.

(c) Pooling of phenotypic descriptors.

Figure 2: Examples of phenotypic descriptors and illustration of the spatial pooling scheme with four interest points.

2.1. Revealing a mimicry-related phenotype with sparse discriminant analysis

The collected phenotypic descriptors are here condensed into a single mimicry-related phenotype. This amounts to determining the low-dimensional manifold, in the high-dimensional feature space, describing the phenotype. We have chosen to use sparse discriminant analysis (SDA) by Clemmensen et al. (2011) for this task. More detail on this procedure can be found in electronic supplementary material ESM 1.

The composite phenotype is constructed as the linear combination $\boldsymbol{\beta}$ of the p descriptors $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_p]$ that best describes the mimicry across the hybridization transect, i.e., the direction in the p-dimensional space that maximizes the ratio of the between-group variance to the within-group variance under elastic net regularization Zou and Hastie (2005).

We define the mimicry-related phenotype for the i'th individual as the projection onto the one-dimensional subspace spanned by $\boldsymbol{\beta}$

$$z_i = \sum_{j=1}^{p} \mathbf{D}_{ij}\boldsymbol{\beta}_j \tag{2}$$

where $\mathbf{D}_{ij}$ is the j'th descriptor value for the i'th individual. For all n individuals this is equivalent to $\mathbf{z} = \mathbf{D}\boldsymbol{\beta}$.

Kernel discriminant analysis (KDA) (Mika and Ratsch, 1999) and Isomap (Tenenbaum et al., 2000) are included as alternative, nonlinear, manifold learning methods. These are further described in the electronic supplementary material ESM 1.

3. A likelihood method for identifying the effective number of genes

We are interested in estimating the effective number of genes, K, affecting a trait, i.e., the number genes required to explain the observed phenotypic variation assuming all genes have the same effect. We assume we have a sample of n individuals from an admixture zone, each with some associated genetic data (e.g., microsatellite data). We will take advantage of the fact that even limited genetic data can be used to infer an admixture fraction for each individual, $\mathbf{f} = \{f_i\}_1^n$, under the assumption that pure forms exist at each end of the transect in the admixture zone. f and $1-f$ then represents the proportion of an individuals genome that is identical to individuals in the right and left end of the transect, respectively. The method we use for estimating the admixture fractions is described in the electronic supplementary material ESM 1, and is based on the kernel discriminant analysis of Mika and Ratsch (1999) with the kernel suggested by Martin (2011). Kernel discriminant analysis allows specification of two known end groups and an explicit scoring of previously unseen individuals in relation to these groups and is thus well suited for this purpose.

We will assume that the phenotypic values, $\mathbf{z} = \{z_i\}_1^n$, are normally distributed, given the underlying genotype, and that each locus contributing to the phenotype has the same effect and dominance factors, and that the effects are additive among loci. We will also assume that each locus is di-allelic and that the allele favoring the phenotype in the right end of the transect has frequency 1 in the right extreme of the transect and frequency 0 in the left end of the transect. We will also, without loss of generality, denote the alleles favoring the phenotype in the right and left ends of the transect by a and A, respectively. An individual with admixture proportion f, assuming independence among the parental contributions, then has genotype AA in any locus with probability $(1-f)^2$.

We consider the phenotype, z, of an individual to be a realization of the stochastic variable Z with the conditional distribution

$$Z \mid \mathbf{g} \sim \mathcal{N}(\mathbf{h}^T\boldsymbol{\mu}, \sigma_e^2) \tag{3}$$

where σ_e^2 is the environmental variance and $\mathbf{g} = \{G_k\}_1^K$ is a vector of the K genotypes

$$G_k = \begin{cases} 0 & \text{if} \quad AA, \quad p(G_k = 0|f) = (1-f)^2 \\ 1 & \text{if} \quad Aa, \quad p(G_k = 1|f) = 2f(1-f) \\ 2 & \text{if} \quad aa, \quad p(G_k = 2|f) = f^2\ . \end{cases} \tag{4}$$

Three averages are used for the conditional Gaussians $\boldsymbol{\mu} = [\mu_0, \mu_1, \mu_2]^T$ and

$$\mathbf{h}_k = [h_0, h_1, h_2]^T \quad \text{where} \quad h_q = \frac{1}{K}\sum_{k=1}^{K} I(G_k == q)$$

i.e., a vector containing fractions of the K genes having the genotypes AA, Aa and aa respectively.

So, for example, if $K = 3$, an individual with genotypes AA, AA and aa in the three loci, respectively, will have mean phenotype $2\mu_0 + \mu_2$.

Thus, in a noise free scenario a single gene would be able to explain a trait as a piecewise constant function (of the admixture proportion) with three steps. K genes would be able to explain a trait attaining $\binom{K+2}{2}$ different values. Here, a noise free scenario would mean no environmental variance in the phenotype *and* no noise caused by the quantification of the phenotype.

To calculate the likelihood, all possible combinations of genotypes must be considered. The set of all possible combinations will be denoted $\mathbf{G}(K) = \{0,1,2\}^K$, i.e., the K'th Cartesian power of possible genotypes, where a single tuple from this set will be denoted $\mathbf{g}_j = [G_{j1}, G_{j2}, \ldots, G_{jK}]$. This set consists of all possible combinations, with replacement, where the order is significant. A total of 3^K such combinations exists.

The probability of a certain combination of genotypes $\mathbf{g}_j$ given the mixture proportion f is

$$p(\mathbf{g}_j|f) = \prod_{k=1}^{K} p(G_{jk}|f)\ .$$

The likelihood of observing the phenotypic trait over the entire population, allowing K genes to contribute to the expression of the trait, is modelled as

$$p_K(\mathbf{z}|\mathbf{f}) = \prod_{i=1}^{n}\left[\sum_{\mathbf{g}_j \in \mathbf{G}(K)} p(z_i|\mathbf{g}_j)p(\mathbf{g}_j|f_i)\right]\ . \tag{5}$$

However, the estimates of f_i may be associated with statistical uncertainty. Ignoring this uncertainty could lead to biased estimates. We therefore provide an alternative formulation that incorporates uncertainty in the estimates of f_i using a bootstrap approach, i.e. we assume that marker loci used for estimation of f_i have been bootstrapped to provide a bootstrap distribution $\{f_i^b\}_{b=1}^B$. The likelihood of observing the phenotypic trait over the entire population, allowing K genes to contribute to the expression of the trait, is then modeled as

$$p_K(\mathbf{z}|\mathbf{f}) = \prod_{i=1}^{n} \left[\sum_{\mathbf{g}_j \in \mathbf{G}(K)} p(z_i|\mathbf{g}_j) \frac{1}{B} \sum_{b=1}^{B} p(\mathbf{g}_j|f_i^b) \right] . \tag{6}$$

For a fixed value of K, we maximize this function for μ_0, μ_1, μ_2, and σ_e^2 using the BFGS algorithm (Fletcher, 1970). We then repeat this procedure for multiple values of K and choose the value of K that maximizes this profile likelihood function as our maximum likelihood estimate of K. To increase the probability of converging to a global maximum we use a scheme with multiple starting points, see electronic supplementary material ESM 2 for details.

We evaluate the performance of the method using simulations allowing for varying heritability and uncertainty in the estimates of f. The heritability is defined as the fraction of the total phenotypic variance V_P that can be attributed to genetic variance

$$H^2 = \frac{V_G}{V_P} = \frac{V_G}{V_G + \sigma_e^2} . \tag{7}$$

The average phenotypic value is $\bar{z} = \sum_{j=1}^{N_G} z_j p_j$ where z_j is the phenotypic value determined by the genotype and p_j is the proportion of individuals with the j'th genotype.

The genetic variance is determined as

$$V_G = \sum_{j=1}^{N_G} \left(z_j - \bar{z}\right)^2 p_j . \tag{8}$$

To simulate data for a phenotype determined by K genes, n mixture proportions $\mathbf{f} = \{f_i\}_1^n$ are drawn, e.g., from a uniform distribution on the interval [0,1]. The genotype for each of the K loci are then drawn from a multinomial distribution with probabilities as in Eq. (4). Phenotypes are then assigned by simulating from a normal distribution as in Equation (3). In simulations with noise in the estimate of f we simulate B samples from a normal distribution with standard deviation σ_f around f_i, such that mixture proportions used for inference $\hat{f}_i^b \sim \mathcal{N}\left(f_i, \sigma_f^2\right)$.

4. Image and microsatellite data

We used published microsatellite data from two sources: Twomey et al. (2013a) (92 samples), Twomey et al. (2014) (36 samples). In addition, we used 157 samples from an unpublished dataset (Twomey et al. in preparation). The final dataset consisted of 285 *R. imitator* individuals from 16 localities in Peru: the 11 localities shown in Figure 1 and 5 localities between Santa Rosa de Chipaota and Achinamisa (i.e., within the banded-striped transition area). For the unpublished microsatellite data, amplification methods follow Twomey et al. (2013a).

We used JPEG compressed images of 6 *R.summersi*, 7 *R. variabilis* and 304 *R. imitator* individuals from the 11 localities shown in Figure 1. The images are 3888×2592 pixels of size captured with a Canon EOS Rebel XS SLR. Both microsatellite data and image data were available for 179 of the *R. imitator* individuals.

5. Results

Phenotypic descriptors. The phenotypic descriptors described in Section 2 were automatically extracted from all 317 images. Different aspects of the patterns in the population are captured by this collection of descriptors, the most dominant being the stripe directionality; for more detail the phenotypic variance captured by these descriptors, see electronic supplementary material ESM 1.

For every individual, the suite of descriptors extracted for the four interest points (left leg, right leg, lower back, upper back) are: Color/non-color ratios for each point of interest, gradient orientation histograms binned in 2 bins at scales $\sigma = [2, 7]$ with tonal range $\beta = 1$ and shape index histograms in 5 bins at scales $\sigma = [4, 8]$ with tonal range $\beta = 1$.

This adds up to a total of $p = 60$ extracted phenotypic descriptors collected in $\mathbf{D} \in \mathbb{R}^{n \times p}$. The columns of this matrix are centered and normalized to unit variance prior to further analysis.

5.1. Mimicry-related phenotype

We use sparse discriminant analysis (SDA) to identify the linear combination of phenotypic descriptors that best captures the variation in mimetic phenotypes. Under the assumption that the mimetic phenotype has been under selection to resemble the phenotypes of either *R. variabilis* in one end of the transect, or *R. summersi* in the other, we use images of seven *R. variabilis* individuals to represent one group and six imaged *R. summersi* the other group, as the training set. The *R. imitator* individuals only enter the analysis to influence the choice of regularization parameter; see details in electronic supplementary material ESM 1.

In the supplementary material, we also provide results when instead using the most extreme *R. imitator* populations, namely those sampled in Sauce and Micaela, to represent the end populations. There are disadvantages and advantages of both of these approaches. Using the model species amounts to defining the mimicry-related phenotype in terms of similarity to those species. This is desirable when the mimicry related phenotype is of prime interest. However, it has the disadvantage that the two model species may differ in traits not mimicked by *R. imitator*. Using the most extreme *R. imitator* populations has the disadvantage that some of the individuals may not be pure mimetic forms. We obtained similar results using either of these approaches, or if we a pool of both the most extreme *R. imitator* populations and the model species individuals (see suppl. information). In the following, we will refer to the extreme groups as the mimicry defining groups, independently of how they were defined. Combinations of these different ways of specifying the mimicry defining groups and the three manifold learning methods used to quantify the phenotype are included as supplementary information in ESM 1.

The mimicry-related phenotypic value for each individual is obtained by projecting onto the direction $\boldsymbol{\beta}$ according to Eq. (2) and will be denoted z_i for the i'th individual. The values are scaled linearly such that the average value for each of the model species is -1 and 1 respectively.

Grouping the individuals by location and ordering them along the transect from south to north (see Figure 1), a boxplot summarizing the mimicry-related phenotypic values as a function of location can be seen in Figure 3.

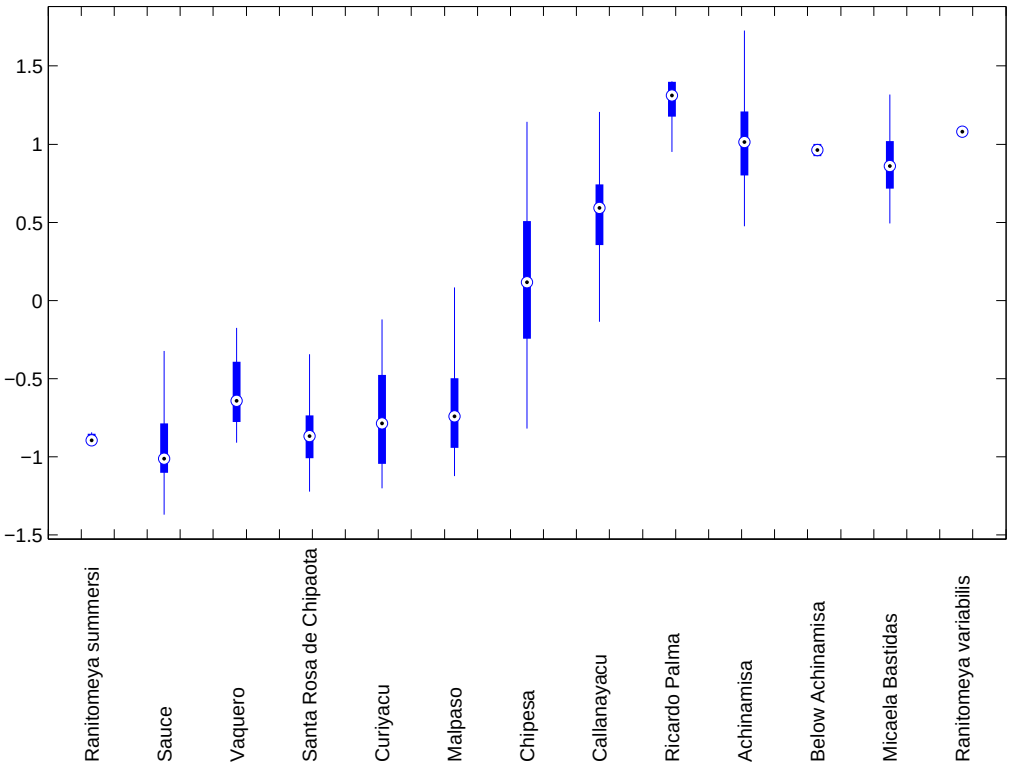


Figure 3: Composite mimicry-related phenotype. Locations are ordered left-to-right from south to north along the Huallaga river. The dot on each box indicates the median, the edges of the box the 25th and 75th percentile and the whiskers extend 1.5 times the inter-quartile range beyond these percentiles.

Notice that the first half of the locations tend to have a value similar to *R. summersi* while the other half have values closer to *R. variabilis*. Chipesa and Callanayacu have phenotypic values that are more intermediate and with relative high variances. Note that the ordering of the locations on the x-axis does not correspond to the actual geographic distances.

5.2. Estimating the number of genes with simulated data

We evaluated the accuracy of the method for determining the number of genes underlying a quantifiable phenotype presented in Section 3 on simulated data for different values of the heritability (see Methods section). The accuracy is evaluated under a variety of scenarios constructed by 1) varying the true number of genes K, 2) sampling the admixture proportions from a uniform or a bimodal distibution, and 3) adding white noise to the admixture proportions. The heritability was varied by simulating data for 1000 different values of σ_e^2. The graphs in Figure 4 show the proportion of runs in which the model assuming $K = 1, 2, 3$ or 4 genes has the highest likelihood.

Generally, the chance of accurate estimation is reduced when 1) the true number of genes is high, 2) the heritability decreases, or 3) the sample size decreases. A measure of confidence in the inference can be obtained by bootstrapping individuals, using the likelihood ratios comparing different hypotheses as statistics. However, if the estimates of f are very noisy, there tend to be a systematic bias towards a higher number of genes for intermediate heritabilities. The effect of this can be seen in Figure 4(d). Using a bootstrap test, we find -6.63 and 0.30 as the 5th and 95th percentiles of the likelihood ratio associated with the null hypothesis of $H_0 : K = 3$ versus $H_1 : K = 4$, for the scenario with a heritability of approx. 0.85, despite the fact that $K = 3$ is the true number of underlying genes. Thus, sensitivity to estimation variance in the admixture proportion must be kept in mind when applying this likelihood model.

5.3. Number of genes underlying the mimicry phenotype

The likelihood model described above was used to estimate the number of genes underlying the quantified phenotype in *R. imitator*. We use 1000 bootstrap replicates to obtain a distribution of likelihood ratios between different alternative models. The bootstrap is performed by sampling individuals with replacement. First a bootstrap distribution of the mixture proportions for each individual is obtained using the available 285 samples. We take into account uncertainty in the estimation of f, by, for each simulation, re-estimating f (see ESM 2) by also bootstrapping microsatellite loci within each individual.

The maximum likelihood values of K, for $K = \{1, \ldots, 5\}$ was then determined for each replicate in a separate bootstrap experiment using the 179 samples with genetic and phenotypic data available. Figure 5 shows (a) a boxplot of the distribution of likelihood ratios associated with the hypothesis $H_0 : K = k$ for $k = 1, 2, 3, 4, 5$, against the alternative hypothesis of $H_A : K = 1$ and (b) the proportion of bootstrap replicates in which each model obtained the highest likelihood value. This proportion can be interpreted as a measure of statistical confidence. In electronic supplementary material ESM 4 the full distribution of likelihood ratios associated with the test of $H_0 : K = 2$ against $H_A : K = 1$ can be seen.

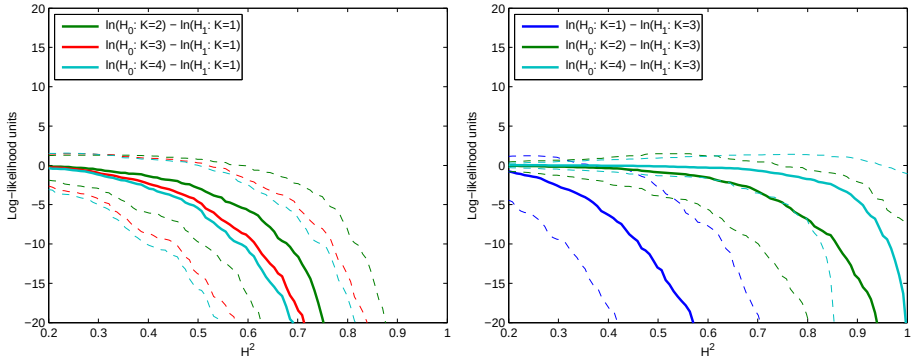
The maximum log-likelihood values are numerically highest for $K = 1$ and in the vast majority of the runs this model is selected as the most likely. The point estimates of the parameters for the hypothesis of $K = 1$ are $[\mu_0, \mu_1, \mu_2, \sigma_e] = [0.882, 0.071, -0.855, 0.274]$.

The p-value associated with different model comparisons are shown in Table 1.

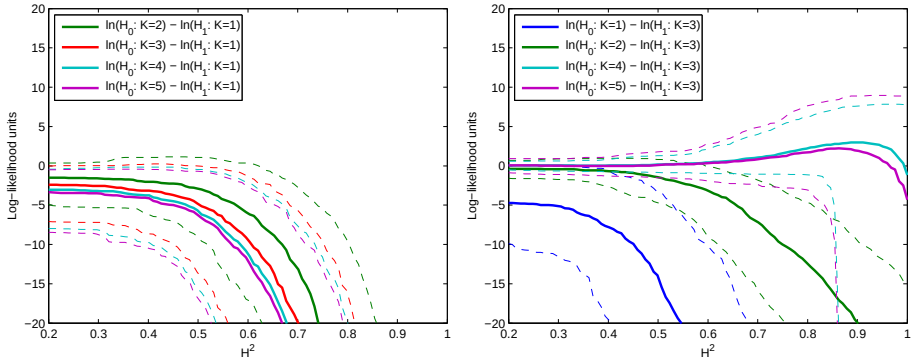
Groups	n_+	$p_{\frac{k=2}{k=1}}$	$p_{\frac{k=3}{k=2}}$	$p_{\frac{k=4}{k=3}}$	$p_{\frac{k=5}{k=4}}$
models	13	0.401	**0.030**	**0.011**	**0.000**
imitator	33	*0.989*	0.136	**0.000**	**0.025**
both	28	*0.964*	0.158	**0.014**	**0.015**

Table 1: P-values for hypotheses of the number of genes, where different mimicry-defining groups are chosen. Regularization parameter $\delta = 0.01$ and number of non-zero loadings is indicated as n_+. Sparse discriminant analysis is used to quantify the mimicry-related phenotype. P-values below 0.05 are typeset in bold and p-values above 0.95 are typeset in italic.

Overall, a model with one ($K = 1$) or two genes ($K = 2$) seems to fit the data best, while three genes ($K = 3$) cannot be rejected. The mimetic phenotype, as measured here is likely mostly influenced by one or two genes of major effect. No combinations of the three different ways of defining the end-populations suggest more than three genes. Using alternative, nonlinear, manifold learning algorithms to quantify the mimicry-related phenotype (see electronic supplementary material ESM 1), only a single combination (KDA with the model species defining the end groups) cannot reject $H_0 : K = 4$ versus the alternative of $H_1 : K = 3$ with a p-value below 0.05.

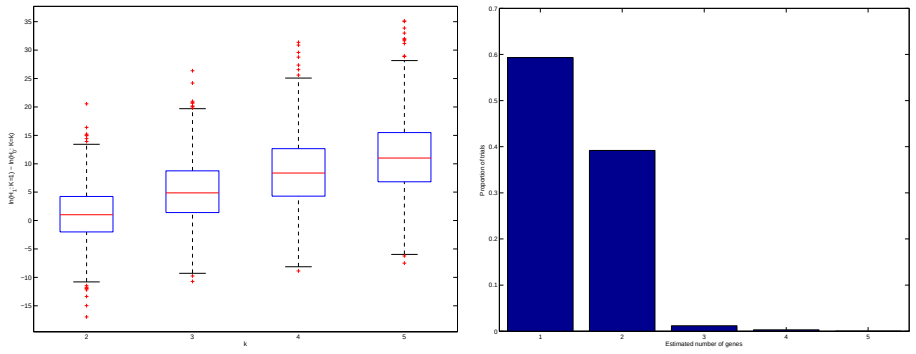


(a) $K = 1, f \sim \mathcal{U}(0,1)$ (b) $K = 3, f \sim \mathcal{U}(0,1)$



(c) $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.01^2)$ (d) $K = 3, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.1^2)$

Figure 4: Likelihood ratios as a function of the heritability H^2 for simulated data. The graphs show median likelihood ratios (solid), 5th and 95th percentiles (dashed) for $K = \{1, \ldots, 5\}$ versus the true K. The captions show the true parameters used to simulate the data for each scenario. 1000 estimations were performed for each of the scenarios. All simulations show that there is never significant support for choosing the wrong value of K, except when the estimation noise on f is high (Figure (d)) where the model is biased toward a higher number of genes for intermediate values of H^2.



(a) Log likelihood ratios $\ln\left(\frac{H_1:K=1}{H_0:K=k}\right)$ (b) Model selection proportions

Figure 5: Maximum log-likelihood, model selection proportions and distribution of differences from bootstrapping 1000 samples. The boxplot summarizes the distribution of maximum log-likelihood values for a model assuming $K = \{1, \ldots, 5\}$ genes. The bar plot shows the proportion of runs where each K has the maximum log-likelihood.

6. Discussion

We have here developed an automated procedure for characterizing complex phenotypes from images. We believe that this method, or related methods, could be of use in many systems where images are available for complex phenotypes. Automated extraction of phenotypic descriptors reduces the subjective biases that may occur when measurements are taken manually and allows for reproducibility of results. While other biases may be introduced through the choice of image capturing system, lighting conditions and/or choice of descriptors, we believe these to be easier to identify and overcome. We notice that such image analyses open up the possibility for a variety of statistical analyses of phenotypes, and their correlations, not pursued here. In this paper, we use the image analyses to define a quantitative measure of the mimetic phenotype in a transition zone between morphs of of *R. imitator*. Using a new method for estimating the effective number of genes affecting this phenotype, we show that the phenotype we measure is likely to be controlled by one or two, or at most three, genes of major effect, and is very unlikely to be affected by many major effect genes. However, there could be substantial phenotypic variation controlled by other genes, but not captured by our quantitative measure of mimetic phenotype.

The fact that we have identified a measure of mimetic phenotype that is controlled by a few genes suggests that future studies aimed at mapping this phenotype have relatively high probability of succeeding. It is substantially easier to map the genes underlying a phenotype controlled by just one or a few genes, than a phenotype controlled by many genes. The phenotype defined here would be useful for such mapping studies.

We can compare our results to *Heliconius* butterflies where the genetic basis of Müllerian mimicry is better understood. In *Heliconius erato*, the transition between the 'postman' and the 'rayed' morphs in the well-studied hybrid zone near Tarapoto, Peru is controlled by three loci of major effect, whereas in the co-mimetic *H. melpomene*, the same mimetic shift (postman to rayed) is controlled by five loci (Mallet et al., 1990). In another example, the polymorphism in *H. cydno alithea* in western Ecuador is controlled by two unlinked loci, one that controls color (white/yellow) and one that controls pattern (presence/absence of melanin in a specific region of the forewing) (Chamberlain et al., 2009). In poison frogs, the genetic basis of color variation is less understood. Early crossing studies in *Oophaga pumilio* (Summers et al., 2004) suggested that pattern is likely controlled by a single locus with a dominant melanin-producing allele, whereas color may be polygenic or controlled by a single locus with incomplete dominance. However, unlike *O. pumilio*, in which a major axis of variation in pattern is presence/absence of melanin, all known populations of *R. imitator* possess melanin on the dorsum, legs, and venter. Thus, a more relevant task in the *R. imitator* system would be identifying the gene or genes that influence the spatial distribution of melanin rather than its presence or absence. Finally, in a field pedigree study (Richards-Zawacki et al., 2012), it was suggested that the red/yellow polymorphism in a population of *O. pumilio* was controlled by a single locus where red coloration was completely dominant over yellow. Thus, our estimates of 1–3 genes controlling the mimetic phenotype of *R. imitator* are fairly comparable to other systems.

The method we have developed for identifying the number of genes controlling a phenotype obtains its information from the degree of clustering of phenotypes and from the dependence of variance in the phenotype on the admixture fraction. It can, as illustrated here, be used to distinguish between a few or many genes, but is not expected to perform well in estimating the exact number of genes, when many genes are involved. In the presence of many genes, the information regarding clustering of phenotypes is lost. We note that the method can be sensitive to the precision in the estimate of the admixture fraction, and results of the method should be interpreted accordingly. Implementations of the presented methods are publicly available at `https://github.com/schackv`.

References

N. Barton. Multilocus clines. *Evolution*, pages 454–471, 1983.

N. Barton and B. O. Bengtsson. The barrier to genetic exchange between hybridising populations. *Heredity*, 57:357, 1986.

N. H. Barton and G. Hewitt. Analysis of hybrid zones. *Annual review of Ecology and Systematics*, 16:113–148, 1985.

D. Briscoe, J. Stephens, and S. O'Brien. Linkage disequilibrium in admixed populations: applications in gene mapping. *Journal of Heredity*, 85(1):59–63, 1994.

J. L. Brown, E. Twomey, A. Amézquita, M. Barbosa de Souza, J. P. Caldwell, S. Lötters, R. Von May, P. R. Melo-Sampaio, D. Mejia-Vargas, and P. Perez-Pena. *A taxonomic revision of the Neotropical poison frog genus Ranitomeya (Amphibia: Dendrobatidae)*. Magnolia Press, 2011.

J. Camper, L. Ruedas, J. Bickham, and J. R. Dixon. The relationship of genome size with developmental rates and reproductive strategies in five families of neotropical bufonoid frogs. *Genetics*, 12:79–87, 1993.

W. Castle. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. *Science*, 54(1393): 223–223, 1921.

R. Chakraborty and K. M. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences*, 85(23):9119–9123, 1988.

N. L. Chamberlain, R. I. Hill, D. D. Kapan, L. E. Gilbert, and M. R. Kronforst. Polymorphic butterfly reveals the missing link in ecological speciation. *Science*, 326(5954):847–850, 2009.

M. Chouteau, K. Summers, V. Morales, and B. Angers. Advergence in müllerian mimicry: the case of the poison dart frogs of northern peru revisited. *Biology letters*, 7(5):796–800, 2011.
C. A. Clarke, P. M. Sheppard, and I. W. Thornton. The genetics of the mimetic butterfly Papilio memnon L. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, pages 37–89, 1968.
L. B. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):37–41, 2011.
J. A. Coyne and H. A. Orr. *Speciation*, volume 37. Sinauer Associates Sunderland, MA, 2004.
J. E. Crawford and R. Nielsen. Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping? *Molecular ecology*, 22 (24):6131–6148, 2013.
J. A. Endler. *Geographic variation, speciation and clines*, volume 10. Princeton University Press, 1977.
R. Fletcher. A new approach to variable metric algorithms. *The computer journal*, 1970.
Z. Gompert and C. Buerkle. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, 18(6):1207–1224, 2009.
C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (...*, 53(2):285–339, 1991.
C. D. Jiggins, R. E. Naisbit, R. L. Coe, and J. Mallet. Reproductive isolation caused by colour pattern mimicry. *Nature*, 411(6835):302–305, 2001.
M. Joron and J. L. Mallet. Diversity in mimicry: paradox or paradigm? *Trends in Ecology & Evolution*, 13(11):461–466, 1998.
J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–564, Oct. 1992.
K. Kunte, W. Zhang, A. Tenger-Trolander, D. Palmer, A. Martin, R. Reed, S. Mullen, and M. Kronforst. doublesex is a mimicry supergene. *Nature*, 507(7491):229–232, 2014.
R. Lande. The minimum number of genes contributing to quantitative variation between and within populations. *Genetics*, 99(3-4):541–553, 1981.
J. Mallet. Hybrid zones of Heliconius butterflies in Panama and the stability and movement of warning colour clines. *Heredity*, 56(2):191–202, 1986.
J. Mallet, N. Barton, G. Lamas, J. Santisteban, M. Muedas, and H. Eeley. Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in Heliconius hybrid zones. *Genetics*, 124(4):921–936, 1990.
F. Martin. An application of kernel methods to variety identification based on SSR markers genetic fingerprinting. *BMC Bioinformatics*, 12 (1):177, 2011.
S. Mika and G. Ratsch. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48, July 1999.
S. P. Otto and C. D. Jones. Detecting the undetected: estimating the total number of loci underlying a quantitative trait. *Genetics*, 156(4): 2093–2107, 2000.
N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. OBrien, and D. Altshuler. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.
K. Pearson. Mathematical contributions to the theory of evolution. XII. On a generalised theory of alternative inheritance, with special reference to Mendel's laws. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 203:53–86, 1904.
C. L. Richards-Zawacki, I. J. Wang, and K. Summers. Mate choice and the genetic basis for colour variation in a polymorphic dart frog: inferences from a wild pedigree. *Molecular ecology*, 21(15):3879–3892, 2012.
M. Slatkin. A method for estimating the effective number of loci affecting a quantitative character. *Theoretical population biology*, 89:44–54, 2013.
K. Summers, T. Cronin, and T. Kennedy. Cross-breeding of distinct color morphs of the strawberry poison frog (Dendrobates pumilio) from the Bocas del Toro Archipelago, Panama. *Journal of Herpetology*, 38(1):1–8, 2004.
R. Symula, R. Schulte, and K. Summers. Molecular phylogenetic evidence for a mimetic radiation in Peruvian poison frogs supports a Müllerian mimicry hypothesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1484):2415–2421, 2001.
R. Symula, R. Schulte, and K. Summers. Molecular systematics and phylogeography of amazonian poison frogs of the genus *Dendrobates*. *Molecular Phylogenetics and Evolution*, 26(3):452–475, 2003.
J. M. Szymura and N. H. Barton. Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland. *Evolution*, pages 1141–1159, 1986.
J. M. Szymura and N. H. Barton. The genetic structure of the hybrid zone between the fire-bellied toads Bombina bombina and B. variegata: comparisons between transects and between loci. *Evolution*, pages 237–261, 1991.
J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319–23, Dec. 2000.
J. R. G. Turner. Two thousand generations of hybridisation in a Heliconius butterfly. *Evolution*, 38:233–243, 1971.
E. Twomey, J. Yeager, and J. Brown. Phenotypic and Genetic Divergence among Poison Frog Populations in a Mimetic Radiation. *PloS One*, 8(2), 2013a.
E. Twomey, J. Yeager, J. L. Brown, V. Morales, M. Cummings, and K. Summers. Phenotypic and genetic divergence among poison frog populations in a mimetic radiation. *PloS one*, 8(2):e55443, 2013b.
E. Twomey, J. S. Vestergaard, and K. Summers. Reproductive isolation related to mimetic divergence in the poison frog Ranitomeya imitator. *accepted for Nature Communications*, 2014.
C. A. Winkler, G. W. Nelson, and M. W. Smith. Admixture mapping comes of age. *Annual review of genomics and human genetics*, 11:65–89, 2010.
S. Wright. *Evolution and the genetics of populations.* University of Chicago Press., 1968.
J. Yeager, J. L. Brown, V. Morales, M. Cummings, and K. Summers. Testing for selection on color and pattern in a mimetic radiation. *Curr Zool*, 58:668–676, 2012.
Z.-B. Zeng. Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: a further improved method. *Genetics*, 131(4):987–1001, 1992.
H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series ...*, 67(2): 301–320, 2005.

ESM 1. Supplementary description of quantitative phenotyping

This supplementary material details important aspects of the methodology used in extracting pattern related phenotypes from images and identifying the mimicry-related manifold in the high-dimensional phenotype space. The general methodology is that a multitude of descriptors are extracted, such that we deliberately extract a surplus of redundant information. In this rich multivariate information, the aspects of the phenotype relating to mimicry will be identified and each individual is assigned a scalar mimicry-related phenotypic quantity. This process is illustrated in Figure 1.

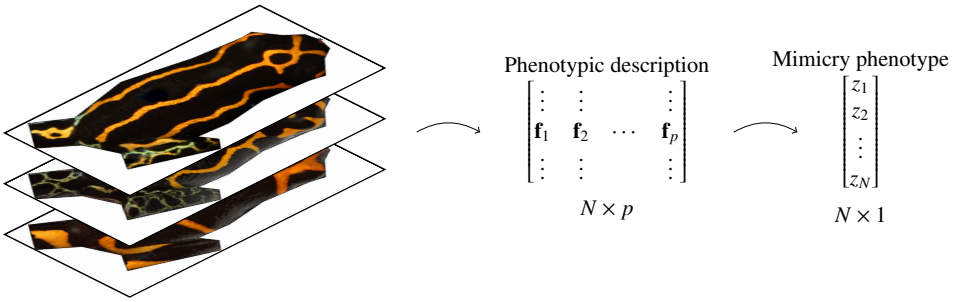


Figure 1: Illustrates the process of image-based quantitative phenotyping in a hybrid zone. A suite of p descriptors are extracted from N images into a $N \times p$ matrix. This matrix contains the entirety of the quantified phenotypic variation for the population. The p dimensions are then reduced to a single dimension, i.e., one scalar per individual, representing the mimicry-related phenotype.

Here we will first elucidate how to extract pattern-related phenotypes from images and get an overview of the phenotypic variation captured. Second, uncovering the relation between the extracted descriptors to the mimicry-related phenotype is detailed. Finally, in section 1.5 we provide supplementary results when using alternative multivariate decomposition methods to identify the mimicry-related phenotype.

1.1. Image-based extraction of pattern variation

The three types of extracted phenotypic descriptors are defined on the pixel-level and described in detail below.

Color/non-color ratio:. A simple descriptor for pattern is a binary value, indicating whether a given pixel is in the black or the colored part of the pattern. To achieve this, the image surface is classified into colored and non-colored (black) regions by stretching the image intensities between its mean ± three standard deviations and thresholding using Otsu's principle (Otsu, 1975).

The color/non-color ratio descriptor's value at image coordinates $\mathbf{x} = [x, y]$ is defined as the value of this binary image $B(\mathbf{x}) \in \{0, 1\}$. Thus, a pooling of this feature will simply be an average of the binary image within the region of the interest point.

Scale space:. The gradient orientation and shape index descriptors below are formulated in a scale space setting (Lindeberg, 1996). This means that they can be extracted at different scales of an image according to the preference of the analyst. Thus the same formulation can be used to extract phenotypic traits independently of the arbitrary scale at which the organism's image was captured.

The scale-space representation of the image $I(\mathbf{x})$ is defined as

$$L(\mathbf{x}, \sigma) = (G * I)(\mathbf{x}; \sigma)$$

where $\mathbf{x} = [x, y]$ and σ is the scale. $(G * I)$ is a convolution of the image with a Gaussian kernel with standard deviation σ. The image derivatives can be calculated in this scale space formulation, where L_x and L_y denotes the gradient in the x- and y-direction respectively and the Hessian matrix $\nabla^2 L = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{bmatrix}$ describes the local curvature. We omit $\mathbf{x}$ from the left hand side of the definitions below for brevity.

Gradient orientation:. The gradient orientation descriptor aims to capture the articulation of stripes on the frog in various orientations.

Gradient magnitude m and orientation θ can be derived from L_x and L_y as

$$m = \sqrt{L_x^2 + L_y^2}\,, \quad \theta = \text{atan2}(L_x, L_y)\,.$$

The gradient orientation is circular on the interval $]-\pi, \pi]$. To quantify the amount of first order change in a given orientation, the gradients are quantized in q bins centered at $b_i, i = 1, \ldots, q$ in this interval.

At a given scale σ, for bin b the gradient orientation descriptor is defined as

$$\text{goh}(b;\sigma) = m(\mathbf{x};\sigma)\frac{\exp\left(\beta^{-2}\cos\left(\theta(\mathbf{x};\sigma) - b\right)\right)}{2\pi I_0(\beta^{-2})} \tag{1}$$

Due to the cyclic nature of the gradient orientations the von Mises aperture is used, where $I_0()$ is the modified Bessel function of order 0 and β is the tonal range. See Larsen (2012) for more details. Note that the gradient orientation contribution is weighted by its magnitude m, which ensures that well defined gradients count more than spurious ones.

Two examples of the signal captured are shown in Figures 2a and 2b, where horizontal and vertical gradients are highlighted on a single scale.

Shape index:. The shape index is a second order image descriptor used to describe local curvature (Koenderink and van Doorn, 1992) and is defined as

$$s = \frac{2}{\pi}\text{atan}\left(\frac{-L_{xx} - L_{yy}}{\sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}}\right), \quad s \in [-1, 1]$$

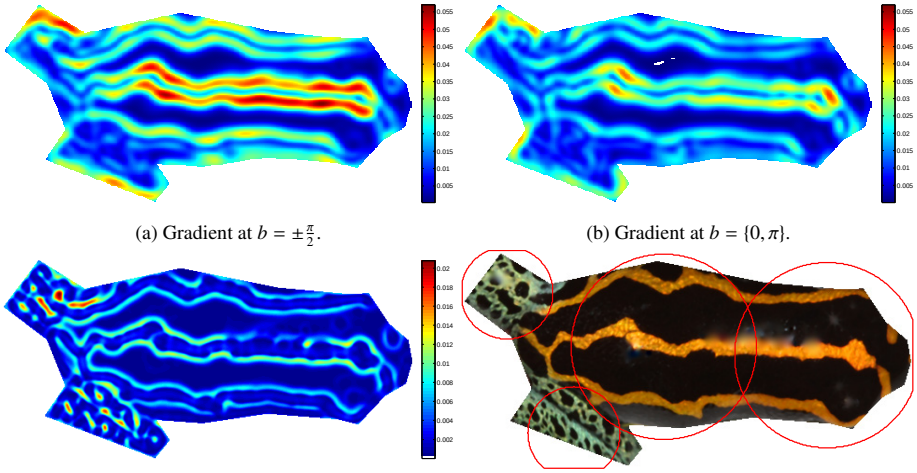
with curvature $c \in \mathbb{R}_+$

$$c = \frac{1}{2}\sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}\,.$$

The binning of the shape index into a histogram is similar to that of the gradient orientation histograms. However, the range of the shape index is not cyclic wherefore a standard Gaussian aperture function can be used. The shape index histogram descriptor for bin b at scale σ is

$$\text{sih}(b;\sigma) = \frac{c(\mathbf{x})}{2\pi\beta^2}\exp\left(-\frac{(s(\mathbf{x}) - b)^2}{2\beta^2}\right) \tag{2}$$

An example of the shape-index response for a single scale and a bin centered at $b = -0.8$ is shown in Figure 2c.



(a) Gradient at $b = \pm\frac{\pi}{2}$.

(b) Gradient at $b = \{0, \pi\}$.

(c) Shape index at $b = -0.8$.

(d) Pooling of phenotypic descriptors.

Figure 2: Examples of phenotypic descriptors and illustration of the spatial pooling scheme with four interest points.

1.2. Phenotypic variance captured

The phenotypic variance captured by the chosen descriptor suite is illustrated here using principal components analysis (PCA) (Jolliffe, 2002). PCA is an eigenvalue decomposition of the correlation (or covariance) matrix which decomposes the multivariate signal into a new coordinate system, where the axes are ordered according to variance explained, i.e., the first principal axis is the axis of maximum variation in the data. The percentage of variance explained by including a given number of principal components is shown in Figure 3. We see that two principal components explains approximately 67% of the variance and that eight principal components collectively explain more than 95% of the variance captured by the phenotypic descriptors. This illustrates that, as expected, there is a significant amount of redundance in the chosen descriptors.

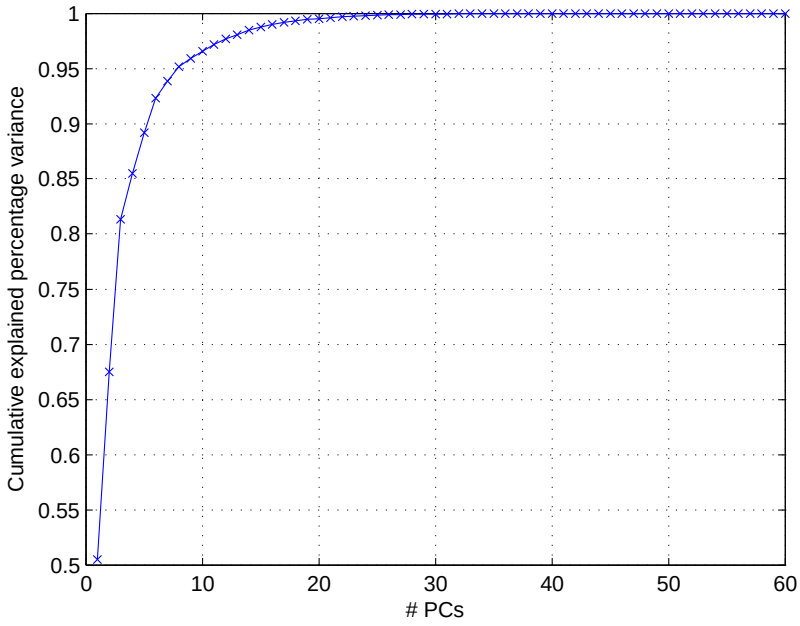


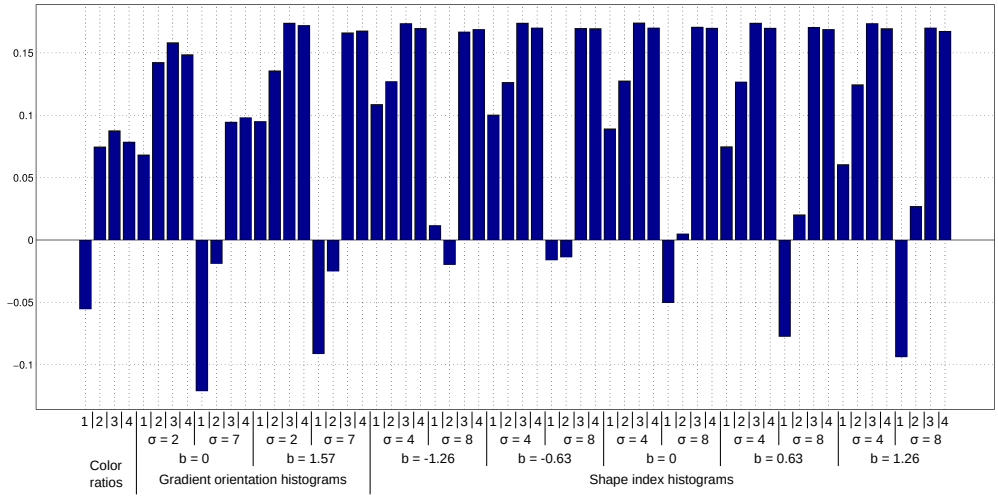
Figure 3: The percentage of variance explained as a function of number of principal components.

To investigate which aspects of the phenotypic variation are captured by each principal component (PC) it is necessary to inspect the loadings. The loadings are the found eigenvectors from the eigenvalue decomposition and carry information on which variables are given weight in each principal component. Figure 4 shows the loadings associated with the first two principal components. The loading vectors are normalized to unit length.

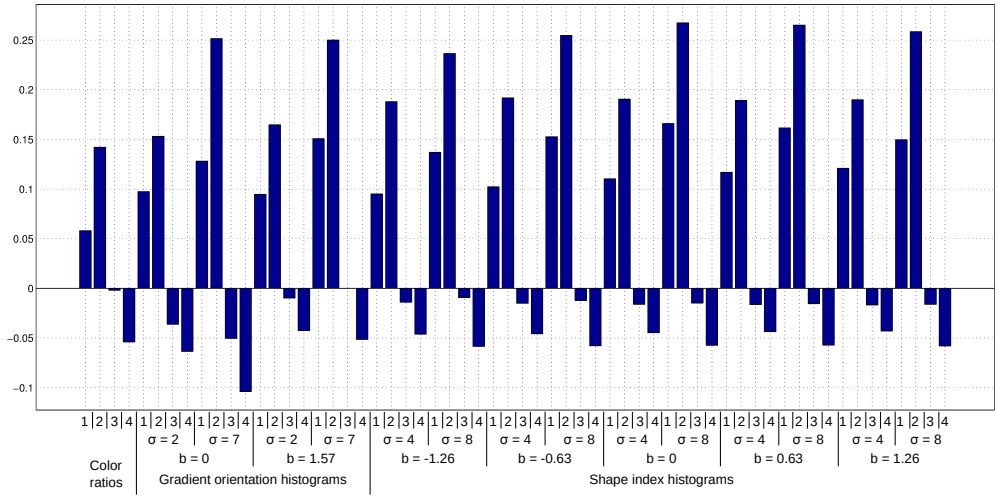
For PC1 all negative loadings are associated with the first two interest points, i.e., the leg patterning. Further we see that the dorsal interest points have very similar loadings. This tells us that the majority of the pattern variation on the two legs are heavily correlated and the same for the dorsal pattern variation. We most importantly observe that not all loadings for the gradient orientation variables are the same; in fact a contrast is evident between horizontal and vertical gradients. This contrast manifests itself in negative loadings for the legs for both vertical and horizontal gradients and positive for the dorsal interest points, and a significant difference in the magnitude of positive loadings for the dorsal interest points between horizontal and vertical gradients.

PC2 is inherently harder to interpret than PC1, since it is constrained to be orthogonal to PC1 which reduces interpretability. However, a clear grouping is still evident: the dorsal interest points are all small or negative, while the interest points for the legs all carry positive loadings. Further we see a difference in the magnitude of positivity between the left and right legs, the left leg carrying the smallest magnitude.

A scatter plot in the coordinate system defined by the first two principal components can be seen in Figure 5. Six individuals are shown as examples on how this principal component space represent the phenotypic variation. Individuals are colored according to the sampling localities as indicated by the legend. Note how vertically banded and horizontally striped frogs are separated along the first principal component. Further note how individuals sampled at each location are loosely clustered along this component. The second principal component is harder



(a) PC1 loadings.



(b) PC2 loadings.

Figure 4: Variable loadings (eigenvectors) for the first and second principal component. From the left are color/non-color ratios, gradient orientation histograms and shape index histograms. The numbers 1 to 4 indicate which interest point the descriptor covers, σ indicates the scale at which the descriptor was extracted and b the bin center. Note how anatomically similar interest points exhibit similarities in these loadings.

to interpret from these examples, but by simultaneous inspection of the loadings it seems differences in left versus right leg patterning and head versus dorsal patterning are accentuated along this axis.

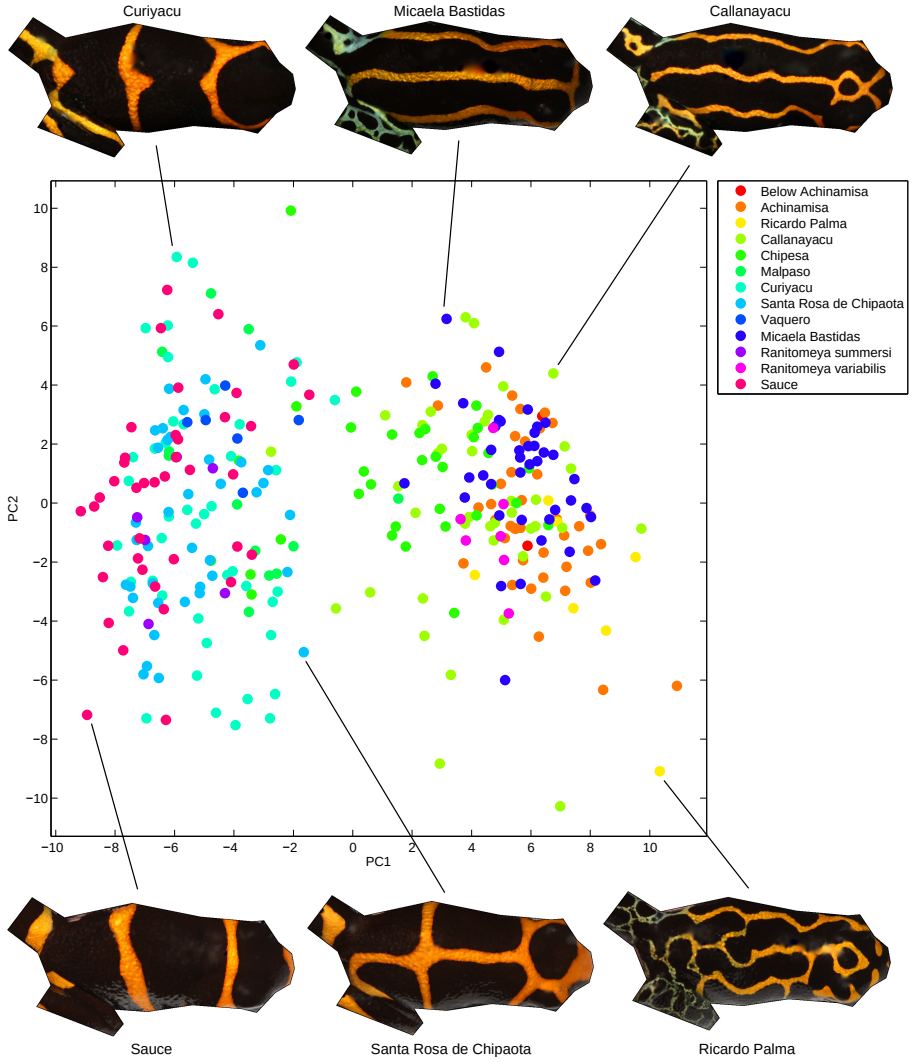


Figure 5: Scatter plot of the first two principal components. Individuals are colored according to sampling location. Six individuals are shown as examples of the phenotypic variation captured by these first two principal components.

We remind the reader that the first principal component account for approximately double the variance of the second principal component. Combined with the orthogonality constraint this makes higher order components inherently harder to interpret. Further, this is a two dimensional projection of a minimum eight dimensional space, wherefore phenotypic characteristics clustering together in this projection, might be separated along higher order components. Therefore such a visualization is mostly useful for identifying very strong signals, such as the stripe directionality in this case.

The phenotypic variation captured by the overcomplete descriptor suite has now been illustrated by using principal components analysis. The next section illustrates how the descriptors relate to the direction in the multivariate descriptor space representing the mimicry-related aspects of this phenotype.

1.3. Identifying mimicry-related variation

The multitude of extracted phenotypic descriptors collectively provide a high-dimensional description of the pattern phenotype. In this high-dimensional space, we seek to identify the one-dimensional mimicry-related manifold, i.e., a single scalar value per individual, representing the degree to which an individual mimic each of the model species. One meaningful way to do this is to let the phenotypic differences of the model species define this one-dimensional manifold, since mimicry-related phenotypic expressions are similar to one model species at one extreme and the other model species at the other extreme. Discriminant analysis techniques are capable of that and a variant called sparse discriminant analysis is the primary method used here (Clemmensen et al., 2011). Other approaches are reviewed in ESM 1.5.

In a two-group scenario, with a total of n_m individuals, SDA recasts the discrimination problem as an optimal scoring problem

$$\underset{\boldsymbol{\beta},\boldsymbol{\theta}}{\arg\min} \; \|\mathbf{Y}\boldsymbol{\theta} - \mathbf{D}_m\boldsymbol{\beta}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{3}$$

$$\text{s.t.} \quad \frac{1}{n_m}\boldsymbol{\theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta} = 1 \tag{4}$$

where $\mathbf{Y}$ is an $n_m \times 2$ dummy matrix encoding the group membership with $Y_{ik} = 1$ if the i'th individual belongs to the k'th group, $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ are the *scores*, $\mathbf{D}_m$ is the $n_m \times p$ matrix of phenotypic descriptors for the individuals belonging to either group and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^T$ is the loadings vector. The scoring vector $\boldsymbol{\theta}$ is constrained to be orthogonal to the trivial solution $\mathbf{1}$ and would in a two class problem, with no regularization, become proportional to the class means. The parameter γ enforces shrinkage, while λ enforces sparsity. Tuning λ is analogous to constraining the maximum number of contributing descriptors. Compared to, e.g., stepwise selection of variables, this formulation takes into account groups of correlated descriptors. This makes SDA especially well suited when the number of extracted descriptors is high compared to the number of individuals in each group.

1.4. Descriptor relation to mimicry

Identifying the extracted descriptors importance for the mimicry-related phenotype can serve the purpose of better understanding the biological system. In the context of SDA, a compact representation of the phenotypic descriptors is achieved by restricting to a maximum of $k \leq p$ non-zero loadings. Note however, that even though only a limited number of descriptors contribute, this does not imply that the other descriptors are uncorrelated with the identified linear combination. Therefore it is meaningful to inspect the correlations between the p original variables and the mimicry-related phenotype, rather than the raw loadings in $\boldsymbol{\beta}$. In Figure 6 these correlations are shown as bars.

A number of the descriptors are strongly positively correlated with the mimicry-related phenotype and a few are negatively correlated. These contrasts are interesting: All of the negative correlations are with descriptors extracted on the legs (1 and 2). The larger of the two scales for the left leg (1) in particular, is negatively correlated. This indicates that the leg patterning exhibits a different type of change across the transect than the rest of the patterning. Further, the gradient orientation descriptors for the dorsal interest points (3 and 4) have strong correlations for horizontal stripes (b=1.57) compared to vertical stripes (b=0). This indicates that the degree of horizontal versus vertical striping partially determines what value of the mimicry-related phenotype an individual will have. The shape-index descriptors exhibit a high correlation with the elements of dorsal pattern, while low (or negative) for leg patterning. The color/non-color ratios show a moderate negative correlation for the first interest point, while the remainder are moderately positive.

Note how similar the correlations are to the loadings for the first principal component in Figure 4. This tells us that the mimicry-related phenotype is similar to the strongest phenotypic signal captured by the chosen descriptor suite.

1.5. Alternative manifold learning methods and groupings

Various methods exist for determining the low-dimensional manifold in the high-dimensional feature space on which the mimicry-related phenotype lies. We have chosen to use sparse discriminant analysis (SDA) because 1) it uses information from the end-populations to determine "what descriptors relate to mimicry?" and 2) the weighting of original components is sparse and thus easy to interpret. Isomap (Tenenbaum et al., 2000) is another well-known method for manifold learning; nearby observations (in feature space) are used to construct a neighborhood graph representing the possibly nonlinear manifold. Finally, a kernel discriminant analysis (KDA) could be used, potentially revealing very nonlinear manifolds through the use of a kernel space formulation (Mika and Ratsch, 1999). See also electronic supplementary material 2. However, with the number of observations available

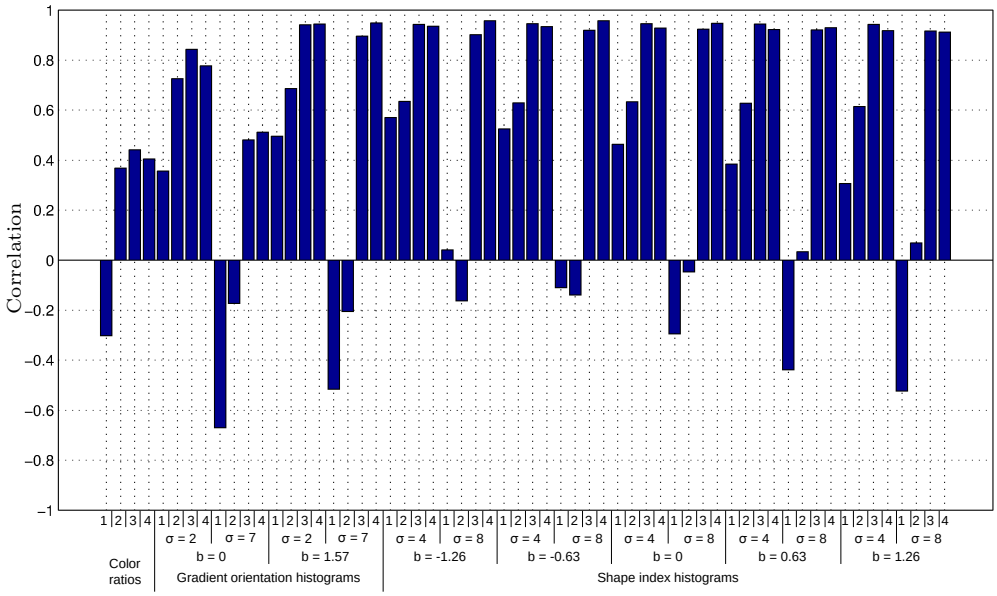


Figure 6: Correlation of phenotypic descriptors with the quantified mimicry-related phenotype. From the left are color/non-color ratios, gradient orientation histograms and shape index histograms. The numbers 1 to 4 indicate which interest point the descriptor covers, σ indicates the scale at which the descriptor was extracted and b the bin center.

for determining the low-dimensional manifold, we find that a linear method is the appropriate choice. It is possible, that with a larger number of samples a more complex manifold could be determined reliably, thus changing these conclusions. Therefore we have included results where each of these alternative methods are used as well.

		Parameters	$p_{\frac{k=2}{k=1}}$	$p_{\frac{k=3}{k=2}}$	$p_{\frac{k=4}{k=3}}$	$p_{\frac{k=5}{k=4}}$
models	SDA	$\delta = 0.010, n_+ = 13$	0.4010	**0.0300**	**0.0110**	**0.0000**
	KDA	$\sigma = 9.540, \lambda = 1.000$	0.0880	0.0700	**0.0330**	**0.0010**
	isomap	$k = 5$	0.2880	**0.0160**	**0.0000**	**0.0000**
imitator	SDA	$\delta = 0.010, n_+ = 33$	*0.9890*	0.1360	**0.0000**	**0.0250**
	KDA	$\sigma = 13.110, \lambda = 1.000$	0.5640	0.4990	**0.0360**	**0.0200**
	isomap	$k = 4$	0.2590	**0.0160**	**0.0110**	**0.0000**
both	SDA	$\delta = 0.010, n_+ = 28$	*0.9640*	0.1580	**0.0140**	**0.0150**
	KDA	$\sigma = 12.430, \lambda = 1.000$	0.5330	0.6940	0.0510	**0.0210**
	isomap	$k = 5$	0.3090	**0.0170**	**0.0030**	**0.0000**

Table 1: P-values for hypotheses of the number of genes, where different methods are used to estimate the low-dimensional manifold in the high dimensional space, describing the mimicry-related phenotype. The leftmost column indicates the mimicry-defining groups. P-values below 0.05 are typeset in bold and p-values above 0.95 are typeset in italic.

Table 1 shows the p-values $p_{\frac{H_0:K=k}{H_1:K=k-1}}$ for rejecting the null hypothesis H_0 of k genes versus the alternative of H_1 $k-1$ genes underlying the mimicry-related phenotype. Three different mimicry-defining group configurations are used, namely using *R. summersi* and *R. variabilis* (models), using *R. imitator* from the two most extreme sampling localities Sauce and Micaela (imitator) or pooling these populations to make up the groupings (both).

For each of the methods, the parameters involved were chosen to fulfil two criteria: a low intra-location variability of the quantified phenotype while still maintaining the proportion of individuals attaining a value between the averages of the two mimicry defining groups. This heuristic is motivated by the fact, that we want to use an as complex a model as possible without collapsing the manifold, i.e., overfitting to the mimicry-defining groups.

The following pages each represent a choice of manifold learning method and grouping, i.e., a row in Table

1. For each configuration three plots are shown: a box plot of the quantified mimicry-related phenotype along the transect, a box plot of log-likelihood ratios and two learning curves used to choose the regularization parameter for the given method. Further, a table of p-values is shown for different combinations of null and alternative hypotheses.

References

L. B. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):37–41, 2011.

I. T. Jolliffe. *Principal component analysis*. Springer New York, 2002.

J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–564, Oct. 1992.

A. B. L. Larsen. An in-depth study of local image descriptors and their performance. Master's thesis, University of Copenhagen, 2012.

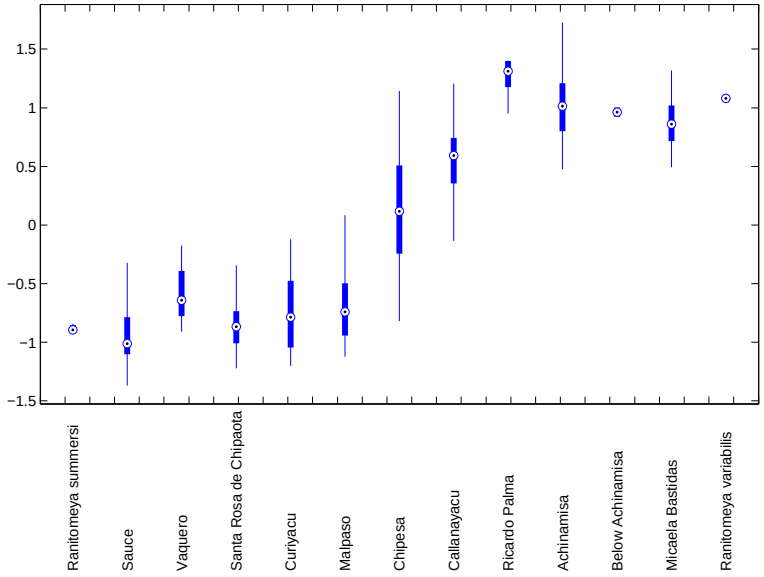
T. Lindeberg. Scale-space: A framework for handling image structures at multiple scales. *CERN European Organization for Nuclear Research - Reports*, pages 27–38, 1996.

S. Mika and G. Ratsch. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48, July 1999.

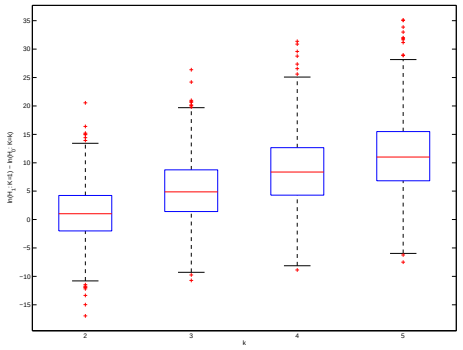
N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23—-27, 1975.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319–23, Dec. 2000.

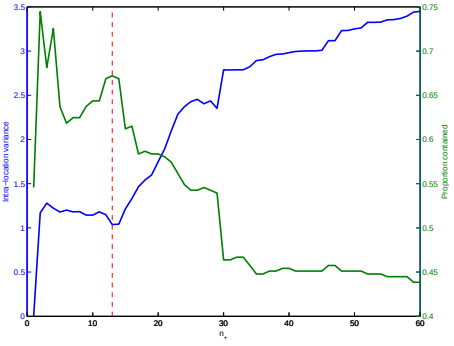
Grouping: models, Manifold method: SDA
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [0.882, 0.071, -0.855, 0.274]$
$K = 2 : [0.958, -0.115, -0.999, 0.235]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

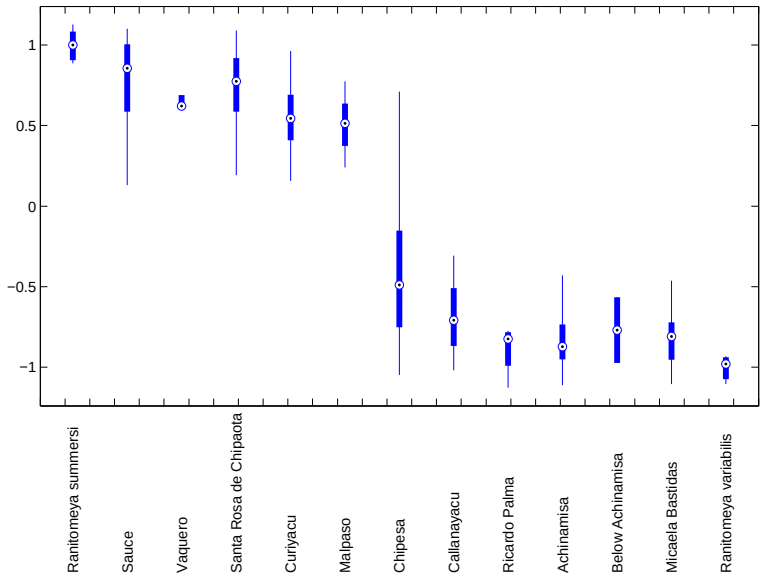


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

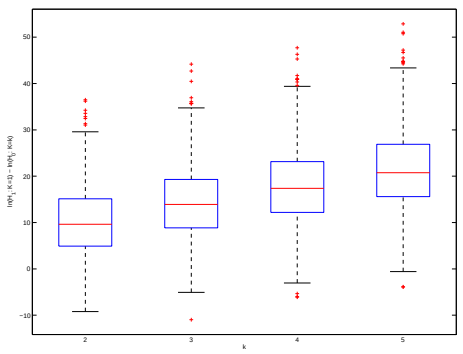
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.401	0.170	0.066	**0.028**
$H_1 : K = 2$	0.599		**0.030**	**0.008**	**0.001**
$H_1 : K = 3$	0.830	0.970		**0.011**	**0.000**

Table 2: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

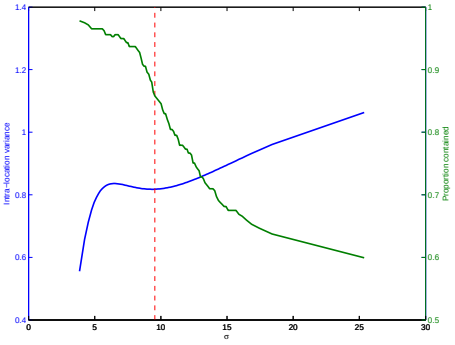
Grouping: models, Manifold method: KDA
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [-0.752, 0.380, 0.710, 0.211]$
$K = 2 : [-0.835, 0.244, 0.781, 0.170]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

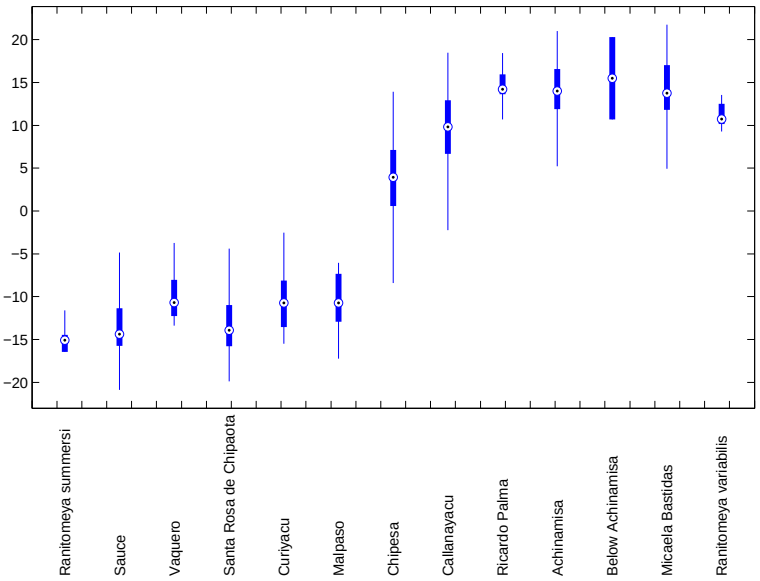


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

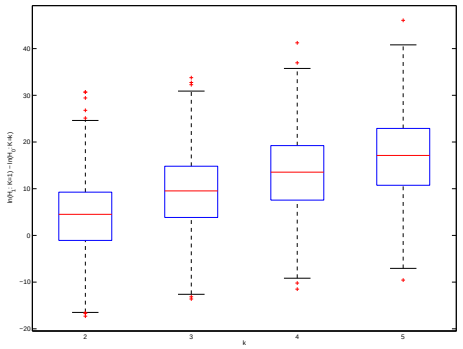
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.088	**0.024**	**0.012**	**0.003**
$H_1 : K = 2$	0.912		0.070	**0.033**	**0.013**
$H_1 : K = 3$	0.976	0.930		**0.033**	**0.007**

Table 3: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

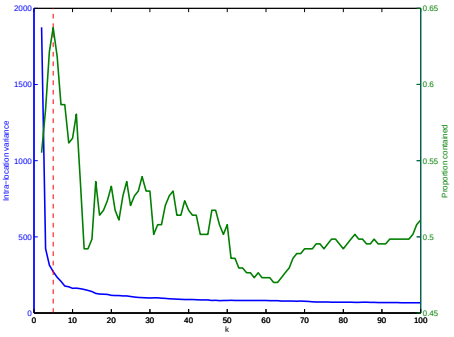
Grouping: models, Manifold method: isomap
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [14.074, 5.794, -11.269, 3.618]$
$K = 2 : [14.497, -6.107, -13.389, 3.226]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

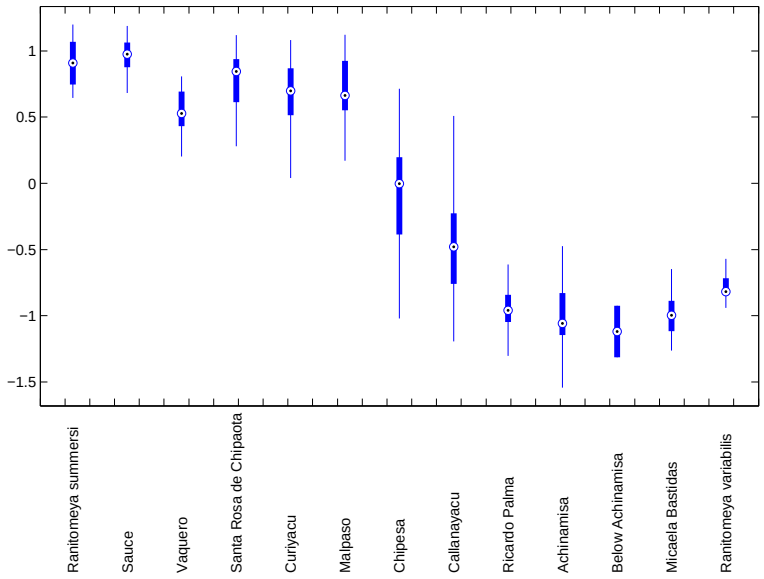


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

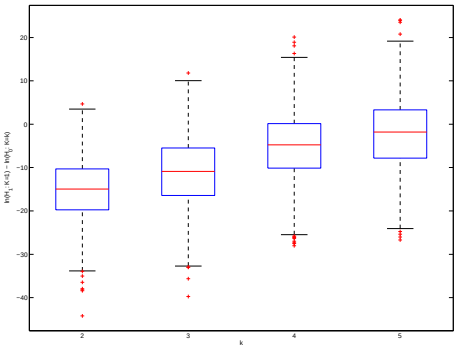
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.288	0.121	0.054	**0.021**
$H_1 : K = 2$	0.712		**0.016**	**0.003**	**0.000**
$H_1 : K = 3$	0.879	0.984		**0.000**	**0.000**

Table 4: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

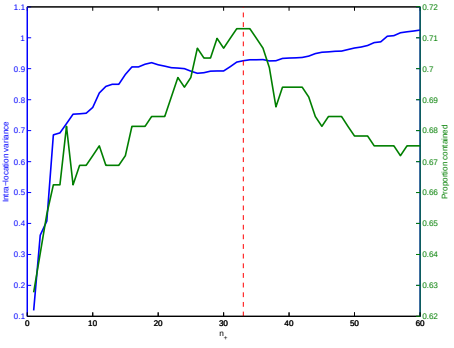
Grouping: imitator, Manifold method: SDA
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [-0.970, -0.120, 0.790, 0.225]$
$K = 2 : [-1.017, 0.147, 0.965, 0.148]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

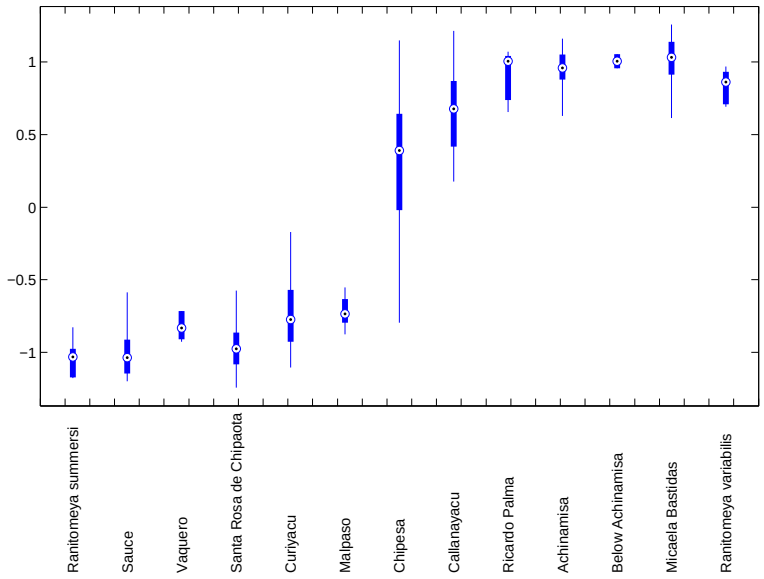


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

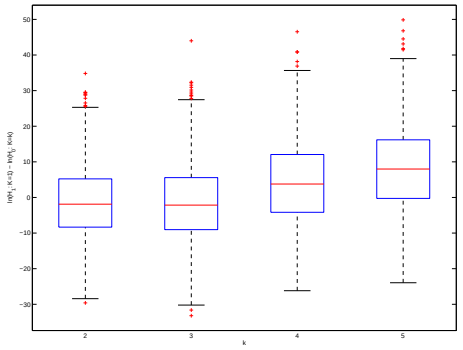
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.989	0.916	0.746	0.601
$H_1 : K = 2$	**0.011**		0.136	**0.005**	**0.004**
$H_1 : K = 3$	0.084	0.864		**0.000**	**0.000**

Table 5: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

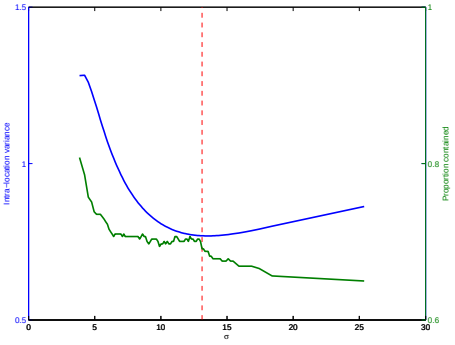
Grouping: imitator, Manifold method: KDA
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [0.977, 0.426, -0.808, 0.213]$
$K = 2 : [0.998, -0.329, -0.998, 0.161]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

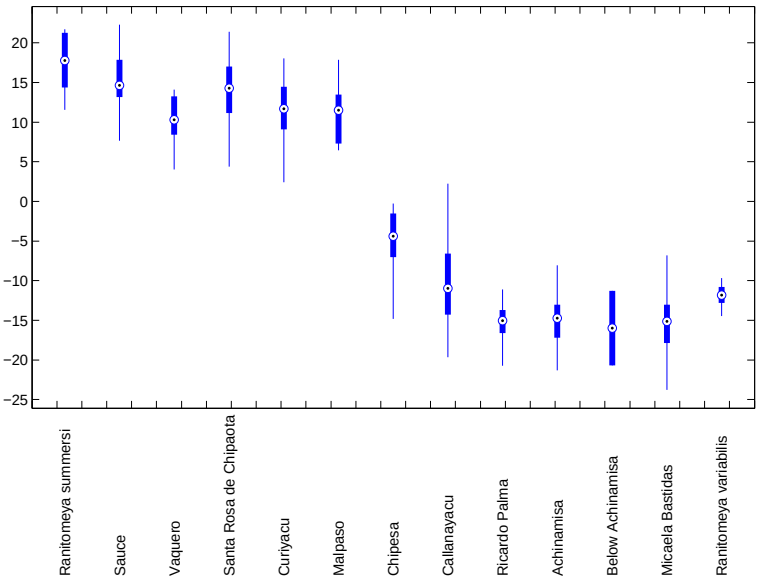


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

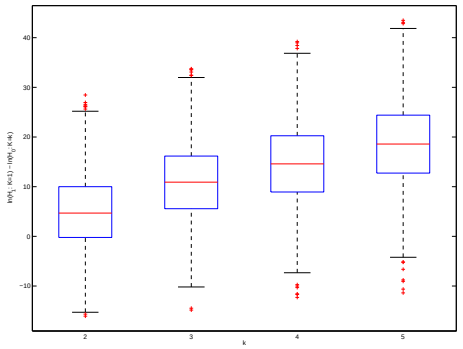
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.564	0.571	0.373	0.260
$H_1 : K = 2$	0.436		0.499	0.117	**0.027**
$H_1 : K = 3$	0.429	0.501		**0.036**	**0.011**

Table 6: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

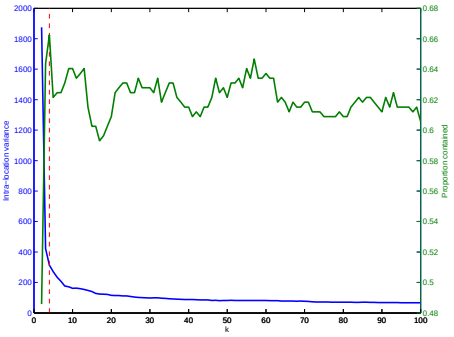
Grouping: imitator, Manifold method: isomap
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [-15.029, -6.009, 12.069, 3.732]$
$K = 2 : [-15.441, 6.673, 14.283, 3.343]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

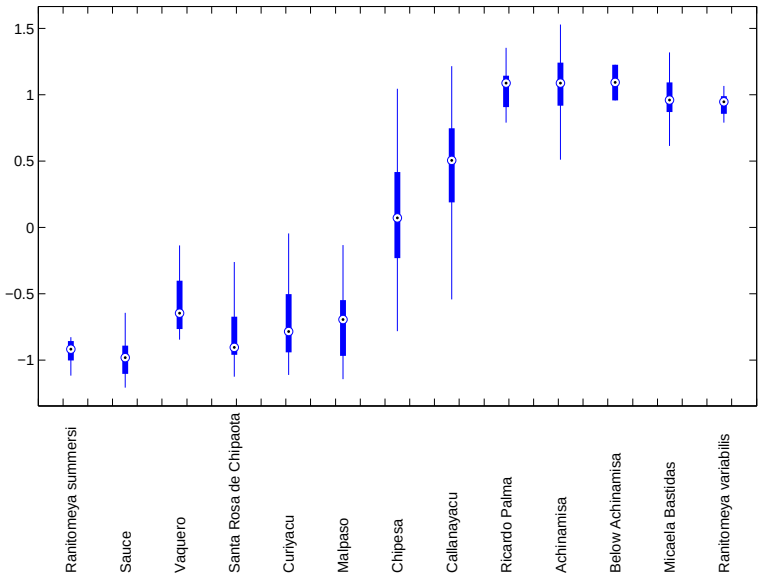


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

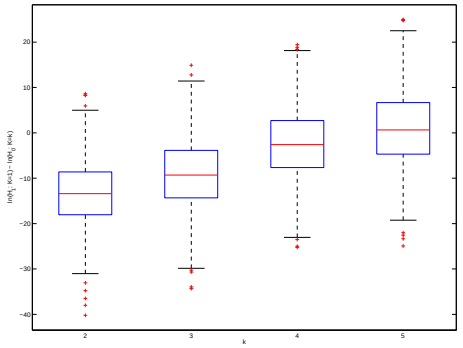
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.259	0.082	**0.048**	**0.030**
$H_1 : K = 2$	0.741		**0.016**	**0.002**	**0.000**
$H_1 : K = 3$	0.918	0.984		**0.011**	**0.001**

Table 7: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

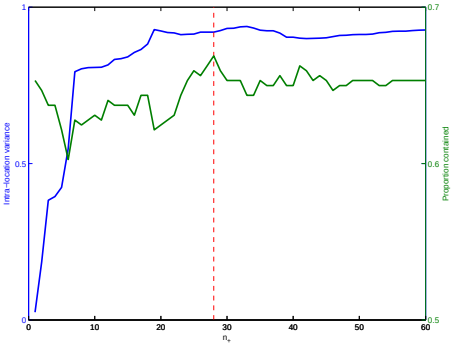
Grouping: both, Manifold method: SDA
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [0.959, 0.150, -0.832, 0.232]$
$K = 2 : [1.018, -0.063, -0.997, 0.151]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

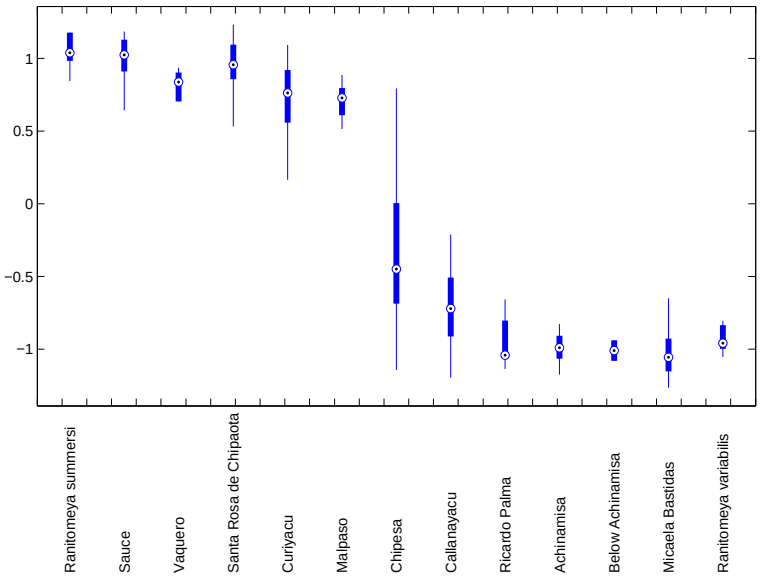


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

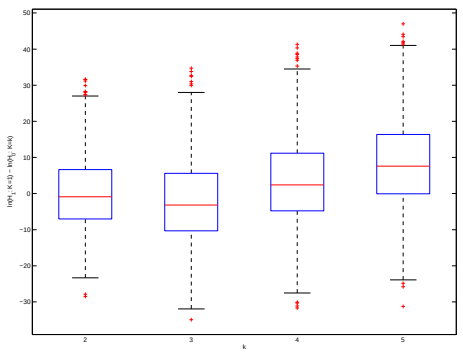
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.964	0.871	0.637	0.474
$H_1 : K = 2$	**0.036**		0.158	**0.012**	**0.004**
$H_1 : K = 3$	0.129	0.842		**0.014**	**0.001**

Table 8: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

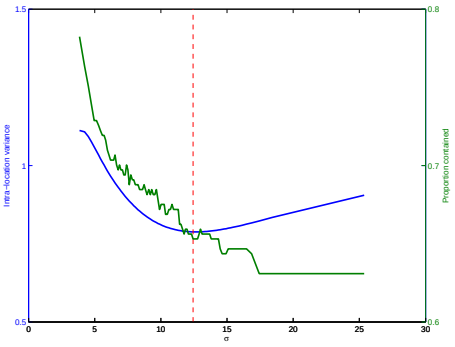
Grouping: both, Manifold method: KDA
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [-0.994, -0.492, 0.797, 0.216]$
$K = 2 : [-1.016, 0.257, 0.992, 0.160]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.

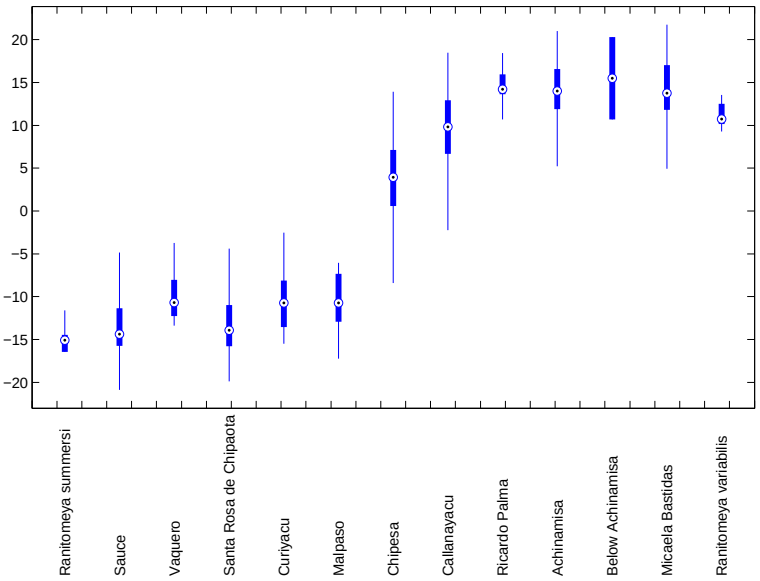


(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

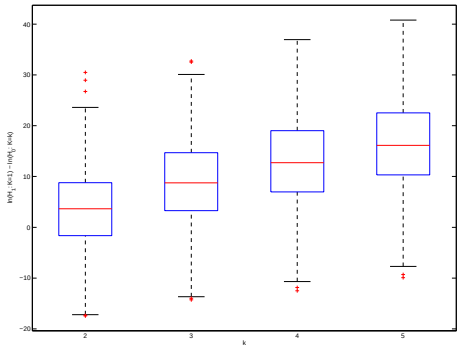
	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.533	0.614	0.418	0.252
$H_1 : K = 2$	0.467		0.694	0.249	0.072
$H_1 : K = 3$	0.386	0.306		0.051	**0.018**

Table 9: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

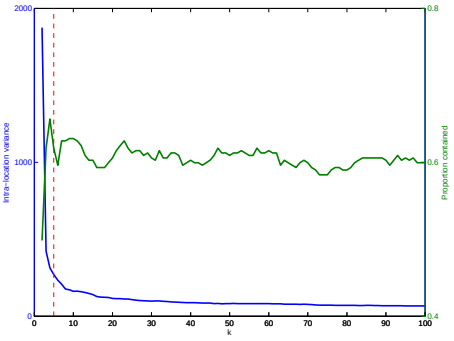
Grouping: both, Manifold method: isomap
Point estimates $[\mu_0, \mu_1, \mu_2, \sigma_e]$:
$K = 1 : [14.069, 5.761, -11.272, 3.616]$
$K = 2 : [14.488, -6.148, -13.387, 3.232]$



(a) Quantified phenotype



(b) Log-likelihood ratios for $H_0 : K = k$ versus $H_1 : K = 1$.



(c) Intra-location variance and proportion samples contained between end group averages as a function of parameter. Dashed red line indicates chosen parameter.

	$H_0 : K = 1$	$H_0 : K = 2$	$H_0 : K = 3$	$H_0 : K = 4$	$H_0 : K = 5$
$H_1 : K = 1$		0.309	0.135	0.068	**0.032**
$H_1 : K = 2$	0.691		**0.017**	**0.001**	**0.000**
$H_1 : K = 3$	0.865	0.983		**0.003**	**0.000**

Table 10: P-values using three different alternative hypotheses $H_1 : K = \{1, 2, 3\}$. Null hypotheses are $H_0 : K = k$ for k being a different number of genes than the alternative hypothesis. P-values below 0.05 are marked in bold.

ESM 2. Supplementary method description

This supplementary material details aspects of the methodology used for optimizing parameters under the proposed likelihood models and estimating admixture proportions from microsatellite data with a data-driven approach.

2.1. Optimization of the likelihood model

Determining the parameters for the global optimum of the likelihood function is crucial in model selection. However, the optimization landscape might be near-flat in some areas or contain local extrema causing the optimization algorithm to converge to non-optimal parameters. To alleviate this, a scheme for choosing starting points and a gradient for maximization of log-likelihood is described here. Both are useful aids in reaching the global optimum rather than a local maximum.

2.1.1. Starting points

The success in finding the global optimum of the proposed log-likelihood function is dependent on the starting point of the optimization algorithm. A scheme with multiple starting points is used here, where the model's conditioning on the genotype is leveraged to select meaningful starting points.

The four parameters $[\mu_0, \mu_1, \mu_2, \sigma]$ are initialized based on simple statistics on the quantified phenotype z. The three averages are initialized at six different points, namely at the following percentiles of z:

μ_0	μ_1	μ_2
10	50	90
10	10	90
10	90	90
40	50	60
40	40	60
40	60	60

These are chosen based on the assumption that a heterozygotic phenotype has a value between the homozygotic phenotypes.

The standard deviation is initialized to fractions of the standard deviation of z, namely $[1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{10}, \frac{1}{15}, \frac{1}{20}]$. This yields a total of 42 starting points, ensuring that the optimization algorithm converges to the global optimum.

2.1.2. Gradient of log-likelihood

We seek the gradient of the log-likelihood with respect to the parameters.

Let $L_K(\boldsymbol{\theta}; \mathbf{z}, \mathbf{f})$ denote the log-likelihood of the parameters $\boldsymbol{\theta} = [\mu_1, \mu_2, \mu_3, \sigma]$ such that

$$L_K(\boldsymbol{\theta}; \mathbf{z}) = \sum_i^n \log \sum_{\mathbf{g}_j \in \mathbf{G}(K)} p(z_i|\mathbf{g}_j) p(\mathbf{g}_j|f_i) \,. \tag{1}$$

Each $\mathbf{g}_j$, $j = \{1, \ldots, M\}$ is represented by the average $\lambda_j = \mathbf{h}_j^T \boldsymbol{\mu}$.

The gradient with respect to each $\mu_t, t = \{1, 2, 3\}$ can be written in terms of the gradient of L_K with respect to $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_M]$. There is one λ_j for every $\mathbf{g}_j$.

The chain rule yields

$$\frac{\partial L_K}{\partial \boldsymbol{\mu}} = \frac{\partial L_K}{\partial \boldsymbol{\lambda}} \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}} \,. \tag{2}$$

The gradient for fixed i, with respect to λ_k, will be derived first.

$$\begin{aligned} \frac{\partial L_K^i}{\partial \lambda_k} &= \frac{\partial}{\partial \lambda_k} \left[\log p(z_i|f_i)\right] \\ &= \frac{1}{p(z_i|f_i)} \frac{\partial}{\partial \lambda_k} p(z_i|f_i) \end{aligned} \tag{3}$$

The genotype probabilities conditional on the admixture proportion can be considered a constant with respect to λ_k:

$$\frac{\partial}{\partial \lambda_k} p(z_i|f_i) = \sum_{j=1}^{M} p(\mathbf{g}_j|f_i) \frac{\partial}{\partial \lambda_k} p(z_i|\mathbf{g}_j) \ . \tag{4}$$

The probability of a phenotype is modeled as a normal distribution with mean λ_j and variance σ^2. For a fixed k the derivative with respect to λ_k is:

$$\begin{aligned}
\frac{\partial}{\partial \lambda_k} p(z_i|\mathbf{g}_j) &= \frac{\partial}{\partial \lambda_k} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(z_i - \lambda_j)^2}{2\sigma^2}\right\} \\
&= p(z_i|\mathbf{g}_j) \frac{\partial}{\partial \lambda_k} \left[-\frac{(z_i - \lambda_j)^2}{2\sigma^2}\right] \\
&= \begin{cases} 0 & j \neq k \\ p(z_i|\mathbf{g}_k)\frac{z_i - \lambda_k}{\sigma^2} & j = k \end{cases}
\end{aligned}$$

Combined with Equations (3) and (4) this yields

$$\frac{\partial L_K}{\partial \lambda_k} = \sum_{i=1}^{n} \frac{p(z_i|\mathbf{g}_k) p(\mathbf{g}_k|f_i)}{\sigma^2 p(z_i|f_i)} (z_i - \lambda_k) \ .$$

Since $\frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}} = \mathbf{H}$ where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M]^T$ the gradient of L_K with respect to $\boldsymbol{\mu}$ is

$$\frac{\partial L_K}{\partial \boldsymbol{\mu}} = \frac{\partial L_K}{\partial \boldsymbol{\lambda}} \mathbf{H} \tag{5}$$

Derivation of the gradient with respect to σ is analogous up until Equation (4). Hereafter

$$\begin{aligned}
\frac{\partial}{\partial \sigma} p(z_i|\mathbf{g}_j) &= \frac{\partial}{\partial \sigma} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(z_i - \lambda_j)^2}{2\sigma^2}\right\}\right] \\
&= -\frac{1}{\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{(z_i - \lambda_j)^2}{2\sigma^2}\right\} \\
&\quad + p(z_i|\mathbf{g}_j) \frac{\partial}{\partial \sigma} \left[-\frac{(z_i - \lambda_j)^2}{2\sigma^2}\right] \\
&= p(z_i|\mathbf{g}_j) \left(-\frac{1}{\sigma} + \frac{(z_i - \lambda_j)^2}{\sigma^3}\right) \ .
\end{aligned}$$

Combining as above yields

$$\frac{\partial L_K}{\partial \sigma} = \sum_{i=1}^{n} \frac{1}{p(z_i|f_i)} \sum_{j=1}^{M} p(z_i|\mathbf{g}_j) p(\mathbf{g}_j|f_i) \left(-\frac{1}{\sigma} + \frac{(z_i - \lambda_j)^2}{\sigma^3}\right) \tag{6}$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \frac{\sum_{j=1}^{M} p(z_i|\mathbf{g}_j) p(\mathbf{g}_j|f_i)(z_i - \lambda_j)^2}{p(z_i|f_i)} \ . \tag{7}$$

Equation (5) gives the first three elements of $\frac{\partial L_K}{\theta}$ and Equation (7) gives the last.

While the derived gradients can seem complicated, the ingredients $p(z_i|\mathbf{g}_j)$, $p(\mathbf{g}_j|f_i)$ are also needed for calculation of the log-likelihood function value (Equation (1)). This is useful for implementation purposes.

2.2. Admixture proportion estimation using kernel discriminant analysis

The admixture proportions are estimated from microsatellite data using kernel discriminant analysis (KDA) (Mika and Ratsch, 1999).

Microsatellite data for a single individual consist of the number of repeats of a given microsatellite at each chromosome. As such, microsatellite data are multi-allelic. It is assumed that a given repeat number occurs as a new mutation in the genome only once. Thus, if two individuals share the same repeat number they are assumed to have inherited it from a common ancestor. If the number of repeats are not the same, the allele is not shared.

Formulating a similarity measure as the proportion of shared alleles opens up microsatellite data to kernel space analysis methods (Martin, 2011).

A brief summary of kernel discriminant analysis is given here.

First, the between-class covariance matrix $\mathbf{M}$ and the within-class covariance $\mathbf{N}$ in kernel space are defined. These are defined in terms of the kernel function $\mathcal{K}(\cdot,\cdot)$. First the j'th element of the mean vector for class i in kernel space is defined as:

$$(\mathbf{m}_i)_j = \frac{1}{\ell_i}\sum_{k=1}^{\ell_i}\mathcal{K}(\mathbf{d}_j, \mathbf{d}_k^i)$$

and then

$$\mathbf{M} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$
$$\mathbf{N} = \mathbf{K}\mathbf{K}^T - \sum_{i=1,2}\ell_i \mathbf{m}_i\mathbf{m}_i^T$$

where

$$\mathbf{K}_{jk} = \mathcal{K}(\mathbf{d}_j, \mathbf{d}_k)$$

Note that $\mathbf{M}$ and $\mathbf{N}$ are here $n \times n$ matrices and $\mathbf{d}_i$ refers to the i'th observation.

Due to the possibility of singularity and the additional need to "capacity control" the feature space, since it can be very non-linear, it is a must to regularize the within-class covariance.

A regularized objective function takes the form as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{M}\mathbf{w}}{\mathbf{w}^T(\mathbf{N} + \lambda\mathbf{I})\mathbf{w}} \quad \lambda \geq 0.$$

This can be solved either as a generalized eigenvalue problem, or, if one is only interested in the direction of the projection vector $\mathbf{w}$, it can be found as $\mathbf{w} = \mathbf{N}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$ (Muller et al., 2001).

It has also been argued that some fraction of the kernel matrix, rather than the identity matrix, could be added for regularization (Nielsen, 2011). This would correspond to penalizing the 2-norm of the projection vector in the original space (Mika et al., 1999). Note once again that $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ is an n-vector, rather than a p-vector as before. Thus the feature space is defined in terms of the n observations used to train the discriminant function.

The projection of a new data point using the kernel discriminant function is less trivial than for the linear method. Due to the fact that the kernel method is formulated in terms of individual-similarities (or inner products), the projection of $\mathbf{d}_{\text{new}}$ takes the form

$$f = \sum_{j=1}^{n} w_j \mathcal{K}(\mathbf{d}_j, \mathbf{d}_{\text{new}})\ . \tag{8}$$

This can be read as a kernelization of the new observation with each of the the training data observation, projected using the discriminating direction $\mathbf{w}$. This implies that the training data set need to be stored for the testing/classification phase.

The value f is the admixture proportion in the context of KDA for microsatellite markers. The groups selected for KDA are the populations of *R. imitator* at each end of the transect (Sauce and Micaela Bastidas). The admixture proportions are linearly such that these two groups have mean 0 and 1 respectively.

References

F. Martin. An application of kernel methods to variety identification based on SSR markers genetic fingerprinting. *BMC Bioinformatics*, 12 (1):177, 2011.

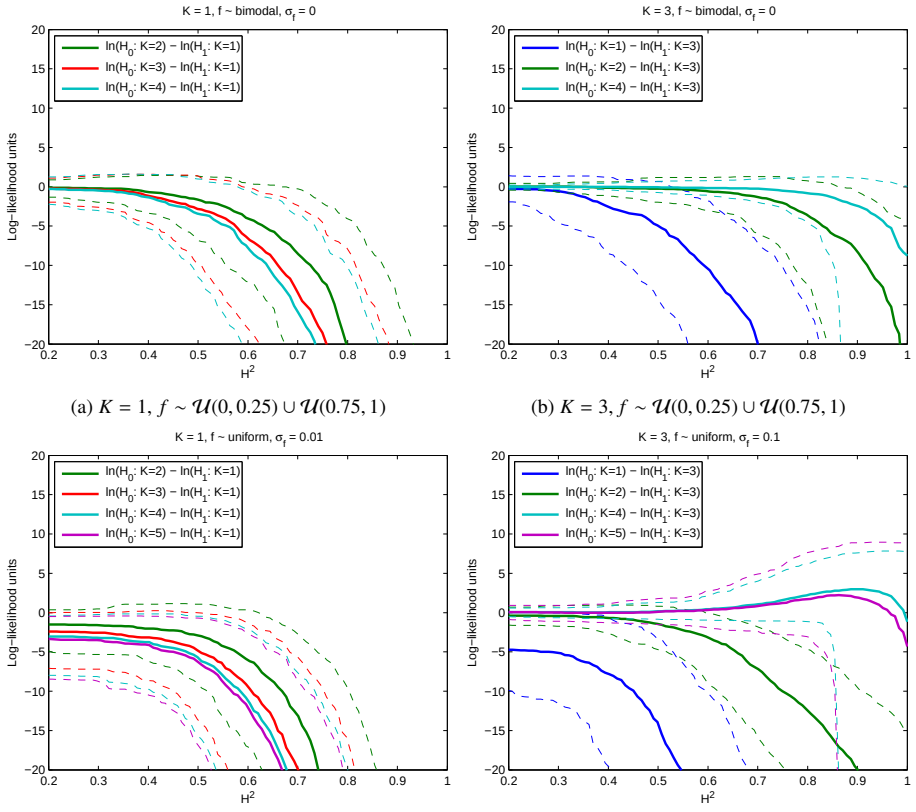
S. Mika and G. Ratsch. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48, July 1999.

S. Mika, G. Rätsch, J. Weston, and B. Schölkopf. Invariant Feature Extraction and Classification in Kernel Spaces. In *NIPS*, volume 89, pages 526—-532, 1999.

K. Muller, S. Mika, and G. Ratsch. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2): 181–202, 2001.

A. A. Nielsen. Kernel Maximum Autocorrelation Factor and Minimum Noise Fraction Transformations. *Image Processing, IEEE Transactions on*, 20(3):612–624, 2011.

ESM 3. Supplementary simulations



(a) $K = 1$, $f \sim \mathcal{U}(0, 0.25) \cup \mathcal{U}(0.75, 1)$

(b) $K = 3$, $f \sim \mathcal{U}(0, 0.25) \cup \mathcal{U}(0.75, 1)$

(c) $K = 1$, $f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.1^2)$

(d) $K = 3$, $f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.01^2)$

Figure 1: Likelihood ratios as a function of the heritability H^2 for simulated data. The graphs show median likelihood ratios (solid), 5th and 95th percentiles (dashed) for $K = \{1, \ldots, 5\}$ versus the true K. The captions show the true parameters used to simulate the data for each scenario. 1000 estimations were performed for each of the scenarios. These simulations are supplements to the main text and include cases where f is simulated as a bimodal distribution.

ESM 4. Bootstrap distribution of log-likelihood ratio

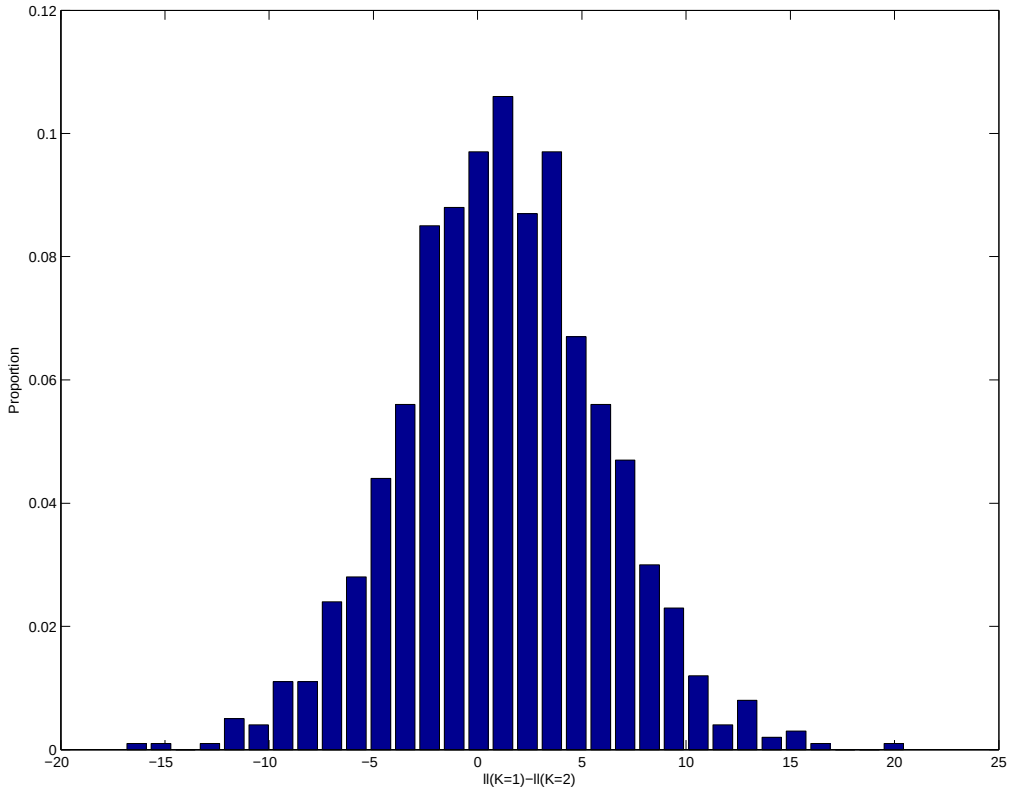


Figure 2: Bootstrap distribution. The histogram shows the distribution of log-likelihood ratios for the null hypothesis of selecting a model with $K = 2$ genes against the alternative of selecting a model with $K = 1$ genes.

ESM 5. Image data

This supplementary material shows all 317 individuals for which image data have been used. The individuals are grouped and ordered according to sampling location as in Figure 3 of the main article.

R. summersi

Sauce

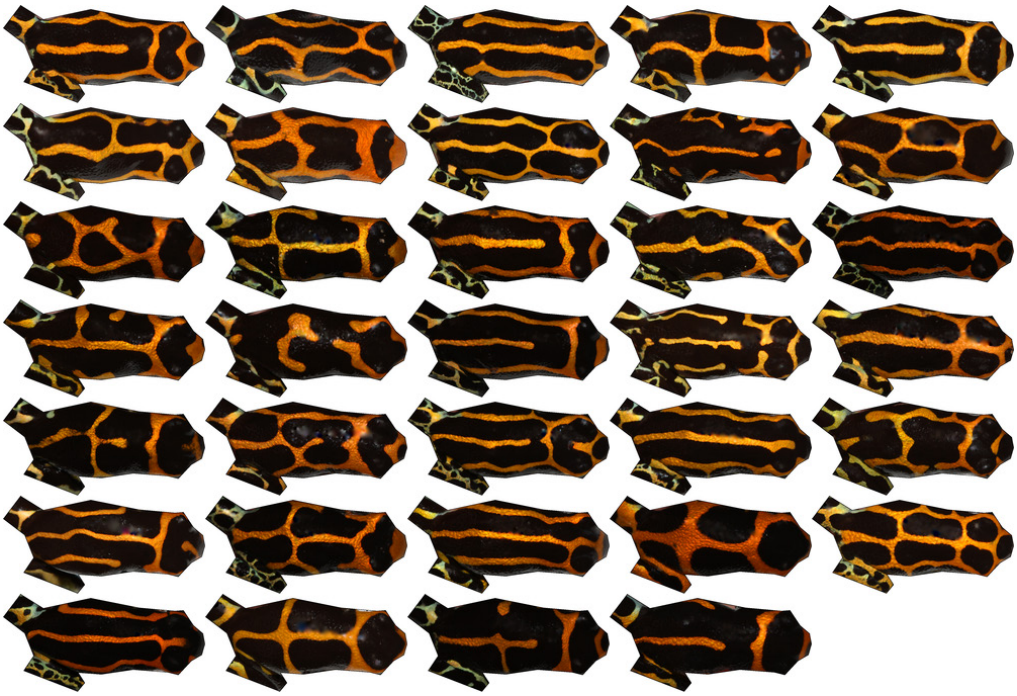
Vaquero

Santa Rosa de Chipaota

Curiyacu

Malpaso

Chipesa

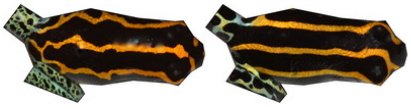


Callanayacu

Ricardo Palma

Achinamisa

Below Achinamisa



Micaela Bastidas

R. variabilis

Paper C

Reproductive isolation related to mimetic divergence in the poison frog Ranitomeya imitator

nature COMMUNICATIONS

ARTICLE

Received 30 Apr 2014 | Accepted 21 Jul 2014 | Published 27 Aug 2014

DOI: 10.1038/ncomms5749

Reproductive isolation related to mimetic divergence in the poison frog *Ranitomeya imitator*

Evan Twomey[1], Jacob S. Vestergaard[2] & Kyle Summers[1]

In a mimetic radiation—when a single species evolves to resemble different model species—mimicry can drive within-species morphological diversification, and, potentially, speciation. While mimetic radiations have occurred in a variety of taxa, their role in speciation remains poorly understood. We study the Peruvian poison frog *Ranitomeya imitator*, a species that has undergone a mimetic radiation into four distinct morphs. Using a combination of colour-pattern analysis, landscape genetics and mate-choice experiments, we show that a mimetic shift in *R. imitator* is associated with a narrow phenotypic transition zone, neutral genetic divergence and assortative mating, suggesting that divergent selection to resemble different model species has led to a breakdown in gene flow between these two populations. These results extend the effects of mimicry on speciation into a vertebrate system and characterize an early stage of speciation where reproductive isolation between mimetic morphs is incomplete but evident.

[1] Department of Biology, East Carolina University, Greenville, North Carolina 27858, USA. [2] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby 2800, Denmark. Correspondence and requests for materials should be addressed to E.T. (email: evan.twomey@gmail.com).

Elucidating the factors that promote population divergence and initiate speciation is key to understanding the evolution of biodiversity. Several studies have identified cases where divergent selection on ecologically relevant traits leads to partial or complete reproductive isolation or speciation[1–4]. Speciation is frequently studied by examining pairs of 'good' species and identifying current reproductive barriers[5]. However, these reproductive barriers may have arisen after speciation was complete, whereas other, currently incomplete barriers may have arisen earlier and been important during initial population divergence[6]. With the goal of investigating initial divergence, one can focus on the early stages of speciation, for example, populations of a single species showing incipient reproductive isolation.

Mimicry can drive phenotypic convergence between distantly related species, but can also drive within-species diversification. This has led to impressive morphological radiations in diverse taxonomic groups such as catfish[7], millipedes[8], snakes[9], bees[10], frogs[11], moths[12] and, most famously, *Heliconius* butterflies[13]. In *Heliconius*, selection for Müllerian mimicry (mimicry between unpalatable species) has led to intraspecific divergence in wing patterns, as different populations radiate into distinct mimicry rings[13]. These wing patterns are also used in mate choice, and morph-based assortative mating can arise as a byproduct of selection for wing mimicry[14] if accompanied by evolution of preferences. Studies of mimetic hybrid zones in *Heliconius* have yielded a range of examples highlighting the continuous nature of speciation. On one end of the continuum, hybrid zones can be narrow and characterized by strong assortative mating, neutral genetic divergence and infrequent hybridization[15,16]. On the other end of the continuum, hybrid zones can be wide, with little or no assortative mating, and with genetic divergence generally restricted to genomic regions controlling colour–pattern differences[17]. There are, however, few examples of 'intermediate' hybrid zones, where distinct mimetic morphs show intermediate levels of genetic divergence and/or premating isolation (but see ref. 18). By identifying cases where speciation appears to have started, but is not yet complete, we can better understand how freely interbreeding populations transition to reproductively isolated species.

Neotropical poison frogs (Dendrobatidae) are diurnal, toxic frogs known for their striking warning colours. A number of species display remarkable intraspecific diversity in colour–pattern[19–22], although in most cases the source of divergent selection among populations is unclear[23–27]. In *Ranitomeya imitator*, intraspecific divergence in colour–pattern is associated with selection for Müllerian mimicry[28], which led to the establishment of four distinct mimetic morphs of this species in central Peru[29]. These morphs resemble three different model species (one of the model species, *R. variabilis*, has two morphs itself[21], both mimicked by *R. imitator*), and occur in different geographic regions, forming a 'mosaic' of mimetic morphs. Where different morphs come into contact, narrow hybrid (or 'transition') zones are formed[29], similar to what has been observed in *Heliconius* butterflies. We have identified three such transition zones, making this study system useful for comparative analyses.

Here we show that a mimetic shift in *R. imitator* is likely driving early-stage reproductive isolation among two of these mimetic morphs. We focus on the narrowest transition zone, which is found in the lowlands of north-central Peru and is formed between the 'varadero' morph, which mimics *R. fantastica*, and the 'striped' morph, which mimics the lowland morph of *R. variabilis*[30] (Fig. 1; Supplementary Fig. 1). Our sampling along a transect crossing this transition zone reveals that there is a shift in several aspects of colour–pattern in *R. imitator*, including dorsal colour (yellow to orange), arm colour (pale greenish-blue to orange), leg colour (pale greenish-blue to navy blue) and dorsal pattern (uniform longitudinal stripes to colouration concentrated around the head). These shifts correspond to the colour–pattern of each model species (Fig. 1; Supplementary Fig. 2), and are therefore likely involved in

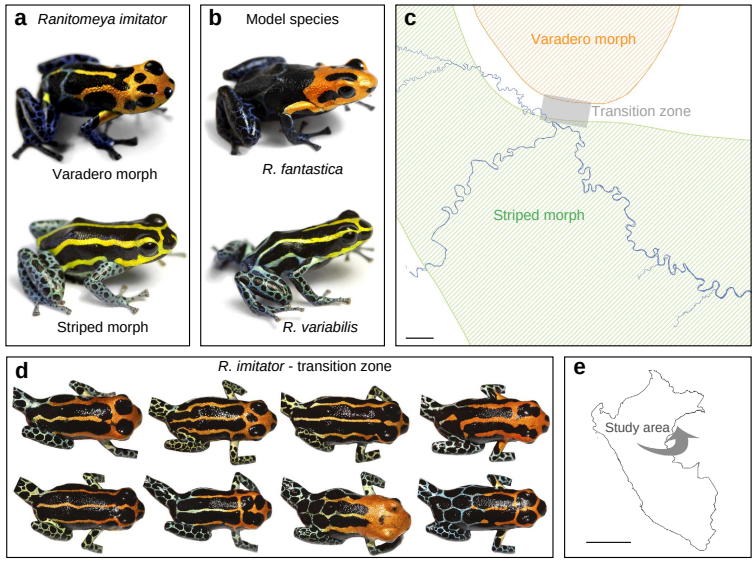


Figure 1 | Mimetic divergence in *R. imitator*. In central Peru, the mimic poison frog *R. imitator* (**a**) exhibits two mimetic morphs corresponding to two different model species (**b**). These morphs occupy distinct geographic areas (**c**), and form a narrow transition zone (grey box, **c**) characterized by phenotypic intermediates (**d**). Scale bar, 3 km (**c**). (**e**) Map of Peru showing study area (scale bar, 500 km).

mimicry. Analyses of colour–pattern clines show that the transition zone is ~1–2 km wide and composed of phenotypic intermediates. Landscape genetic analyses indicate that neutral genetic divergence between morphs is primarily associated with divergence in mimetic colour–pattern, rather than geographic distance, suggesting that mimetic divergence has reduced gene flow between morphs. Using mate-choice experiments, we find evidence for assortative mating in one of the mimetic morphs, however, this mating preference is only present near the transition zone, consistent with reproductive character displacement (RCD). Taken together, these results suggest that mimetic divergence in *R. imitator* has led to a breakdown in gene flow between these two populations, potentially facilitated by assortative mating.

Results

Colour–pattern clines. Selection for different mimetic morphs across geographical areas should cause differentiation in mimetic traits. At the interface between distinct mimetic morphs, traits subject to divergent selection are expected to show a sigmoidal pattern of variation across this zone of mixing[31]. To quantify colour–pattern variation along the mimicry transect, we used a combination of spectrometry and computer-automated feature extraction to extract six colour–pattern variables in *R. imitator*. Transect variation in three of these colour–pattern variables (head colour, body colour and leg pattern) was best described by a linear model (Fig. 2), suggesting gradual spatial change. However, due to our sampling pattern, we cannot rule out the possibility of a sigmoidal cline with a displaced centre for these colour–pattern variables. The remaining three colour–pattern variables (arm colour, leg colour and body pattern) were best described by a sigmoidal model (Fig. 2), suggesting that these aspects of the colour–pattern are under divergent selection. If multiple aspects of the colour–pattern are involved in mimetic resemblance, then shifts in traits should coincide geographically. We tested for cline coincidence among arm colour, leg colour and body pattern by comparing Akaike weights (w_i) between two models: one where cline centre is constrained to a single parameter shared across all three data sets and one where centre is unshared. A common centre was found for all three colour–pattern variables without a significant reduction in model fit (w_i shared centre model = 0.841; w_i unshared centre model = 0.159), indicating coincidence among the three colour–pattern clines. The point estimate for the shared centre parameter was 0.54 km (that is, 0.54 km N from the *a priori*-estimated centre), corresponding to 1.25 km N from the village of Varadero (Supplementary Fig. 1). An alternative explanation for coincident clines is recent secondary contact between divergent populations (see below for discussion of primary versus secondary contact).

In a tension zone model[31], where divergent selection is opposed by dispersal, the width of the cline reflects a balance of selection (which narrows a cline) and dispersal (which widens a cline). Differences in cline widths among different traits may be due to differences in the strength of selection on loci underlying those traits. If different traits are controlled by the same number

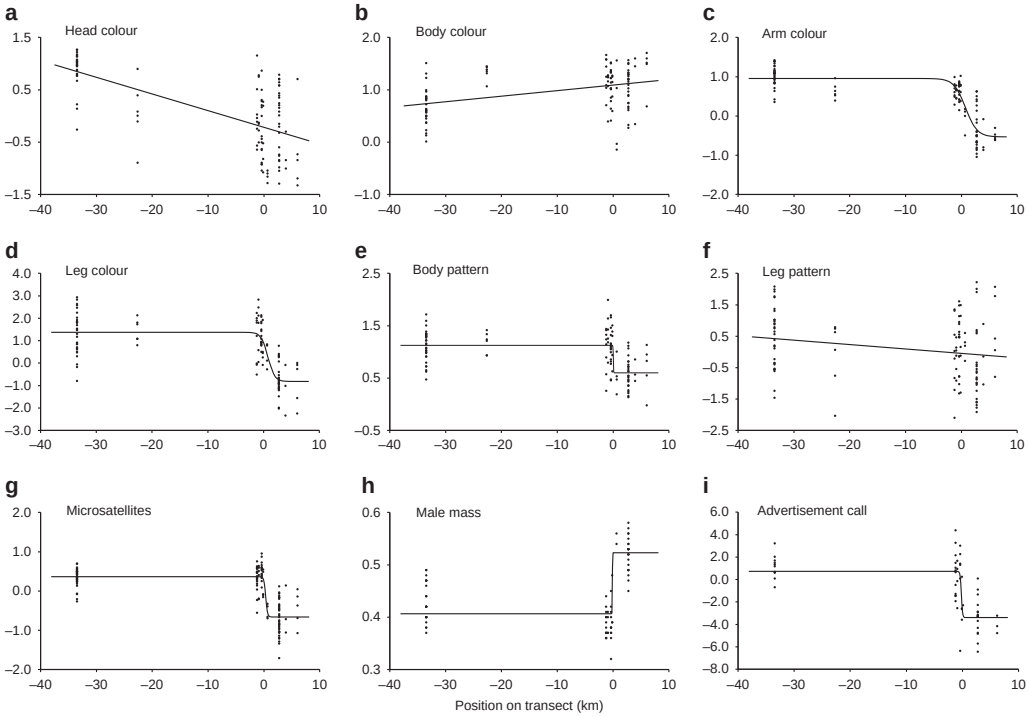


Figure 2 | Clines in colour–pattern, microsatellites, male mass and advertisement calls. In all panels, trait values for individual *R. imitator* (represented by dots) are plotted along the geographic transect (*x* axis). (**a**–**f**) Colour–pattern variation (*y* axis: kernel discriminant score; values closer to +1 indicate closer similarity to *R. variabilis* and closer to −1 *R. fantastica*); (**g**) microsatellite variation (*y* axis: first major axis from factorial correspondence analysis (FCA)); (**h**) male mass (*y* axis: grams); and (**i**) advertisement call variation (*y* axis: linear discriminant score). The fit line for each variable represents the best-supported model describing transect variation and parameter point estimates.

of loci, those under stronger selection should show narrower clines than those under weaker selection. We tested for a common cline width (concordance) among all six colour–pattern variables. A common width could not be found among all six variables without a reduction in model fit (w_i shared width model = 0.273; w_i unshared width model = 0.727), indicating that some colour–pattern variables show non-concordant widths. This was due to the inclusion of the three non-sigmoidal variables, as it was possible to fit a common width of 2.27 km among the three variables showing a sigmoidal pattern of variation (w_i shared width model = 0.747; w_i unshared width model = 0.253). Cline width should be primarily a function of selection strength (assuming constant dispersal), so the evidence that these three clines can be constrained to a common width suggests equivalent strength of selection on arm colour, leg colour and body pattern. This could also suggest a common genetic basis or linkage among all three traits, although colour and pattern elements in dendrobatids are likely controlled by different genes[32].

Landscape genetics. Reduced gene flow between adaptively diverged populations (isolation by adaptation; IBA) is a key prediction of ecological speciation[1]. This results in a positive correlation between adaptive ecological divergence and genetic differentiation among populations after controlling for the effect of isolation by distance (IBD)[33]. Results from the Structure analysis (Fig. 3a) indicate the presence of three genetic groups within the study area. One of these groups (Fig. 3b) is associated with an allopatric population, while two of the groups (Fig. 3c,d) form a sharp break at the mimetic transition zone. These latter two groups still show some evidence of genetic exchange, as there were a few individuals with a striped colour–pattern but a varadero genotype, and vice-versa (Fig. 3). The narrow genetic cline is also characterized by a peak in linkage disequilibrium (Supplementary Fig. 3), further suggesting a barrier to gene flow among the two mimetic morphs. The coincidence of genetic clines and colour–pattern clines was supported by a factorial correspondence analysis, where the cline centre on the first major axis (0.31 km) is almost identical to the shared colour–pattern cline centre (0.54 km), supporting the hypothesis that a shift in mimicry has led to a breakdown in gene flow among mimetic morphs. Using a causal modelling framework, the best-supported hypothesis was one where colour–pattern distance (IBA), but not geographic distance (IBD), was correlated with genetic distance among populations. Multiple-matrix regression[34] yielded similar results, except that both colour–pattern distance ($r^2 = 0.427$, $P = 0.006$) and geographic distance ($r^2 = 0.230$, $P = 0.001$) were accounted as significant predictors of genetic distance. However, the correlation coefficient for colour–pattern distance is nearly twice that of geographic distance, indicating that IBA is a stronger determinant of among-population genetic divergence than is IBD. An alternative interpretation for these results is mimetic divergence in allopatry followed by secondary contact. This could explain the neutral genetic divergence among these two populations, however, one would expect the microsatellite cline to be wider than the observed 0.54 km unless contact happened very recently (see below).

Mate-choice experiments. One potential mechanism for a breakdown of gene flow between adaptively diverged populations is morph-based assortative mating[35]. To address the role of assortative mating, we conducted triad mate-choice experiments in which we introduced two females (one of each morphs) into the terrarium of a given male, and measured the amount of courtship time between the male and female. This is equivalent to a mutual choice test, which is appropriate here as *R. imitator* is monogamous[36], and therefore both sexes should be choosy. We tested preferences in three populations: striped allopatric, striped transition and varadero, allowing us to address two questions (1) whether courtship preferences differ between the striped-transition and varadero populations, and (2) whether courtship preferences differ among the two populations of the striped morph. Using generalized linear mixed models (GLMM), we found an overall significant effect of male origin ($\chi^2_{1,55} = 16.518$, $P = 0.00026$), indicating that mate preferences were significantly different across populations. A *post hoc* test revealed that the preferences in the striped-allopatric and varadero populations

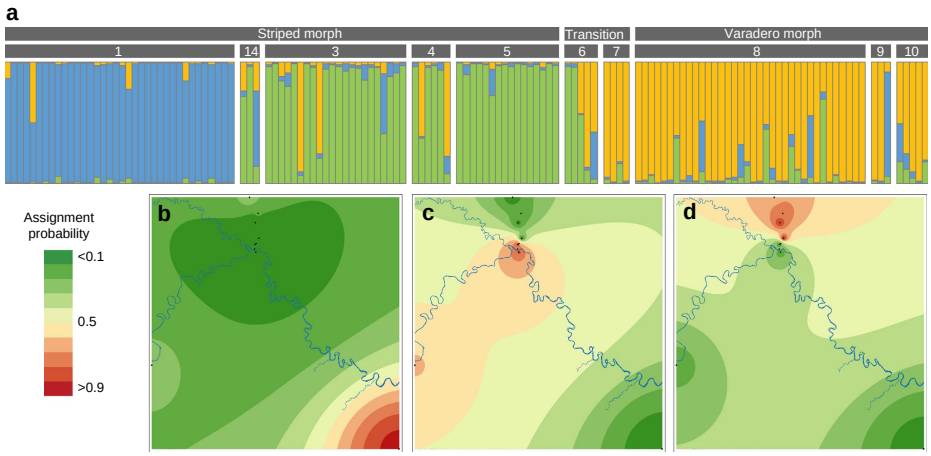


Figure 3 | Genetic structure between mimetic morphs of *R. imitator*. (**a**) We used the software Structure 2.3.4 to analyse the multilocus microsatellite data set and assign individuals of *R. imitator* to each of *K* populations. The optimal number of inferred populations was $K = 3$ (shown). Vertical bars indicate membership fractions to inferred groups 1 (blue), 2 (green), and 3 (orange). Horizontal grey bars represent the morph (upper bar) and sampling localities (lower bar). (**b**–**d**) Spatial genetic structure of each of the three genetic groups as inferred by Structure. We projected the Structure output to a map by interpolating the average probability assignment score of each population to each inferred group using inverse-distance-weighted interpolation in ArcGIS. (**b**) Probability assignment to group 1, (**c**) probability assignment to group 2 and (**d**) probability assignment to group 3.

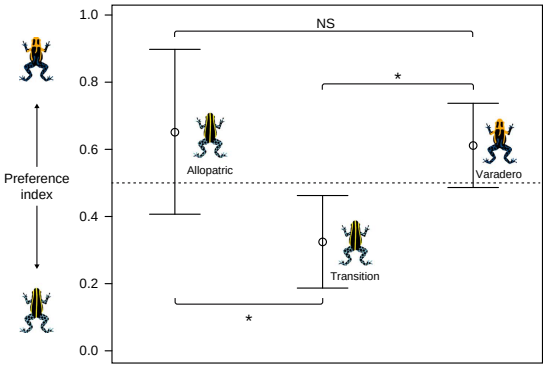


Figure 4 | Courtship preferences in *R. imitator*. For display on the figure, raw courtship times for each trial were converted to a 'preference index', which was calculated by dividing the time a male spent courting the varadero female by the time spent courting the varadero female + time spent courting the striped female (that is, dividing by total courtship time). This index therefore ranges from 0 (all courtship with striped female) to 1 (all courtship with varadero female), with a value of 0.5 (indicated by the dotted line) indicating no preference. Open circles show the mean preference index for each population; error bars represent 95% confidence intervals. Icons next to the error bars represent the morph of the male used in the experiment. Asterisks indicate significant differences from the GLMM ($P<0.05$) from the *post hoc* tests, following FDR adjustment for multiple comparisons. Sample sizes are as follows: striped allopatric, $N=10$; striped transition, $N=19$; varadero, $N=26$.

were not significantly different ($\chi^2_{1,22}=3.096$, false discovery rate (FDR)-adjusted $P=0.078$), with neither population showing a significant preference (Fig. 4). However, preferences between the striped-transition and varadero populations were significantly different ($\chi^2_{1,33}=11.986$, FDR-adjusted $P=0.00161$), mainly due to the striped-transition population's preference towards its own morph (Fig. 4), which indicates that mating preferences have diverged between these two populations across the transition zone. Finally, preferences between the striped-allopatric and striped-transition populations were significantly different ($\chi^2_{1,29}=9.748$, FDR-adjusted $P=0.00269$), suggesting that mating preferences in the striped-transition population are stronger at the mimetic transition zone.

Bioacoustics. During our sampling, there were apparent differences in the advertisement calls of the striped and varadero morphs, which could represent a potential premating isolating mechanism between the two morphs. To determine whether the pattern of call variation coincided with the mimetic transition zone, we recorded the calls of *R. imitator* across the sampling transect. The call of *R. imitator* is a short, musical trill of 0.44–1.07 s, with trills (or 'notes') repeated roughly every 4–20 s, and a dominant frequency of 4,710–5,660 Hz. Each note is composed of 16–32 pulses, with an average pulse rate of 24–30 pulses per second[21]. Note length was negatively correlated with temperature ($r^2=0.115$, $P=0.006$), and pulse rate was positively correlated with temperature ($r^2=0.259$, $P<0.001$). To account for this, we standardized each of the three bioacoustic variables by calculating regression residuals against temperature. After temperature standardization, two bioacoustic variables showed a sigmoidal rather than linear pattern of variation across the transect (note length: Akaike weight (w_i) linear model = 0.032, w_i sigmoidal model = 0.967; dominant frequency: w_i linear model = 0.006; w_i sigmoidal model = 0.992). The point estimates of cline centre were similar (note length centre = −0.14 km; dominant frequency centre = −0.41 km), indicating that the shift in these two call parameters occurs in roughly the same geographic location. Furthermore, the estimated cline centres both occur very close to the estimated colour–pattern and microsatellite cline centres (within <1 km), indicating that the shift in call characteristics occurs in the same place as the shift in colour–pattern and microsatellites. For pulse rate, the linear model was favoured (w_i linear model = 0.830, w_i sigmoidal model = 0.138), indicating a smooth, rather than abrupt, transition across the putative transition zone. To derive a single metric-describing call variation, we used a linear discriminant analysis to derive a discriminant score where the two groups for classification were defined as the populations on the end points of the transect (that is, populations 1 and 10 in Supplementary Table 1). Both note length and dominant frequency contributed substantially to the discriminant function, whereas pulse rate did not (standardized canonical discriminant function coefficients: dominant frequency = 1.343; note length = −1.189; pulse rate = 0.162). This metric showed a sigmoidal pattern of variation with similar cline centre and width as observed in the colour–pattern metrics (Fig. 2; Supplementary Table 2).

Discussion

Our mate-choice trials found that the preferences in the two striped populations we studied were stronger in the striped-transition zone population relative to the striped-allopatric population, a pattern consistent with RCD. However, our experimental design is limited in terms of inferring RCD given that we only tested three populations, and therefore, assuming that mating preferences vary among populations, there is a one in three chance that the strongest preference will be in striped-transition population. A much more robust test of RCD would involve testing multiple populations to determine whether contact among morphs explains variation in mating preferences. Patterns of enhanced mating preferences in areas of contact have, however, been observed in mimetic *Heliconius* butterflies. For example, *H. melpomene* populations that are sympatric with *H. cydno* display stronger mating preferences relative to allopatric populations[37]. In another example, mating preferences in both *H. cydno* and *H. pachinus* are much stronger in sympatry than allopatry[38]. One explanation for this pattern is reinforcement, where mate preferences are strengthened in zones of sympatry to avoid producing unfit hybrids. However, several other processes can result in a pattern of enhanced mate preferences in zones of sympatry (for example, differential fusion hypothesis and 'noisy neighbour' hypothesis; see ref. 6 for review). In this case, as non-mimetic hybrids may suffer fitness costs if they experience higher predation rates[39], adaptations to avoid cross-morph matings are expected to be favoured by selection.

In addition to a shift in colour–pattern and microsatellites at the transition zone, we found a shift in body mass and certain aspects of the advertisement call. Striped frogs south of the transition zone tend to have a smaller body size and a shorter, more highly pitched call compared with the varadero morph north of the transition zone. For both body size and advertisement calls, variation along the transect is best described by a sigmoidal cline with centres coinciding with the colour–pattern and genetic clines (Supplementary Table 4), further supporting the existence of a transition zone. This also supports the possibility for secondary contact, where clines are expected to be congruent for multiple traits. One possible explanation for the shift in body size is that *R. variabilis*, the model species of the

smaller, striped morph of *R. imitator*, is smaller than *R. fantastica*, the model species of the varadero morph (*R. variabilis* mass: $\bar{x} = 0.52$ g, $n = 3$; *R. fantastica* mass: $\bar{x} = 0.68$ g, $n = 4$). Thus, size could represent a mimetic adaptation. As our experiments did not address the specific cue used in mate choice, the roles of colour–pattern, body size and advertisement calls in mediating mate choice in this system should be investigated further.

As we have mentioned above, secondary contact among differentially adapted populations could give rise to many of the observed cline patterns. One plausible scenario here would be mimetic divergence in allopatry, followed by secondary contact. Determining whether hybrid zones are the result of primary or secondary contact without historical evidence is difficult[31]. However, secondary contact with neutral diffusion is unlikely given our dispersal estimate in *R. imitator* of 0.095 km per generation (see Supplementary Methods for details on dispersal calculations). The cline created by secondary contact with subsequent neutral diffusion would exceed the observed overall cline width (0.97 km, see Supplementary Table 4, model D) in only 17 generations, or ~11 years. Considering secondary contact, a more likely scenario is that the cline is maintained by some isolating barrier. In either case (primary or secondary contact), the cline is associated with a shift in mimicry, and may be maintained, at least in part, by assortative mating. Overall, the existence of a narrow cline, as well as moderate genetic divergence between morphs (F_{ST} between mimetic morphs is 0.065–0.077), suggests that mimetic divergence may be playing a key role driving early-stage speciation in a vertebrate system.

Methods

Data availability. Colour–pattern data, advertisement call data, mate-choice data and the full microsatellite data set are available at Dryad (doi:10.5061/dryad.rd586).

Sample collection and transect description. For colour–pattern analyses, we sampled a total of 127 *R. imitator* from 15 localities in the department of Loreto, Peru. Ten of these localities (localities 1–10 in Fig. 1) lie on a rough north-south transect 40 km in length, running from the village of Micaela Bastidas in the south to 7 km N from the village of San Gabriel de Varadero in the north. We sampled an additional five localities off the transect but still relevant for inferring the spatial arrangement of the two focal morphs of *R. imitator*. For genetic analyses, we sampled 136 *R. imitator* from 10 localities. Tissue samples for genetic analysis (toe clips) were taken with sterile surgical scissors and preserved in 96% ethanol before extraction. In most cases, both tissue samples and colour–pattern measurements were taken from each frog, although there were some localities where only genetic data were collected or only colour–pattern data were collected (see Supplementary Table 1 for details). In addition, we took colour–pattern measurements from the two putative model species: 7 *R. variabilis* from Pongo de Cainarachi (representative of the typical lowland *R. variabilis* morph) and 7 *R. fantastica* collected from San Gabriel de Varadero.

Because the transect is not perfectly linear, we calculated transect position as straight-line distance from the putative transition zone centre, with localities south of this point given a negative sign and localities north of this point given a positive sign. The initial centre point (latitude/longitude: $-5.70653°$, $-76.41427°$) used in these calculations was estimated from field observations where an apparent shift in colour–pattern occurred. Therefore, instances where the estimated cline centre from nonlinear regression was close to zero indicate a close fit to our field observations. Cline centre estimates with a negative sign indicate the inferred cline centre to be south of the initial centre point, whereas positive values indicate the cline centre to be north of the initial centre point.

Colour and pattern quantification. To quantify frog colour, we measured the spectral reflectance at specific points on the dorsal surfaces of the mimic species (*R. imitator*) and both model species (*R. variabilis* and *R. fantastica*). Two measurements were taken on the head (right and left sides), four on the body (right and left sides of mid-body and rump) and two on the legs (dorsal surface of right and left thighs). Reflectance measurements were taken using an Ocean Optics USB4000 spectrometer with an LS-1 tungsten–halogen light source and Ocean Optics SpectraSuite software. A black plastic tip was used on the end of the probe so that measurements were always taken at a distance of 3 mm from the skin and at a 45° angle. White standards were measured for every other frog using an Ocean Optics WS-1-SL white reflectance standard to account for lamp drift. Spectral data were then processed in Avicol version 6 software[40] using Endler's segment model[41] calculated between 450–700 nm. This model calculates brightness (Qt), chroma (C), hue (H) and two Euclidean coordinates representing position in a two-dimensional colour space: blue-yellow axis position (MS) and red-green axis position (LM). Measurements within body regions (head, body and legs) were averaged. In addition to the spectrometer measurements, we measured upper-arm colouration using the colour-picker tool (set to a 5×5 pixel average) in Adobe Photoshop CS4 on dorsal photos of each frog, recording the average intensities of red, green and blue channels on two points on each upper arm. Photos were taken on a white background using a Canon Rebel XS DSLR with a Canon EF 100 mm macro lens and the camera flash.

We quantified frog pattern by a collection of local image descriptors. The descriptors were automatically extracted from images of every individual and collected in a feature matrix. Three types of descriptors were extracted: a colour/non-colour ratio, gradient-orientation histograms and shape-index histograms[42–44]. Collectively, these capture zeroth-, first- and second-order image structure. A spatial pooling scheme was used to separately collect information at four interest points: left leg, right leg, lower and upper back. At each of these interest points, pattern variation occurs on a distinct scale, wherefore the descriptors were extracted according to a scale-space formulation[45]. Colour/non-colour ratios were extracted for every interest point on a single scale; gradient-orientation histograms for every interest point on two different scales and two orientation bins (horizontal and vertical), and shape-index histograms were only extracted for the legs on two scales in five bins equidistantly spaced between $-\pi/2$ and $\pi/2$. This summed up to a total of $4 \cdot (1+2 \cdot 2+2 \cdot 5) = 60$ features per individual.

To reduce the multivariate colour and pattern data to a single descriptive metric per body region, we used kernel discriminant analysis[46], where the two model species (*R. variabilis* and *R. fantastica*) represented the training groups used for classification. This procedure assigns a discriminant score to each *R. imitator* individual on the basis of their similarity to either model species, and thus can be thought of as a 'mimicry score'. The analysis can be constrained to include only subsets of the variables to derive a metric for different body regions, for example, leg colour variation in *R. imitator*. Kernel-based analysis is implicitly capable of estimating nonlinear effects, making it more suitable for non-normally distributed features, such as the colour metrics output from Avicol. Using this procedure, we derived colour metrics for four body regions (head, body, legs and arm) and pattern metrics for two body regions (dorsum and legs). For additional details on kernel discriminant analysis, see Supplementary Methods, Supplementary Table 5, and Supplementary Figs 4 and 5.

Cline analysis. To describe clinal variation in colour–pattern elements (as well as average male mass, advertisement call and microsatellites; see Supplementary Methods), and in particular to estimate cline width, we performed nonlinear regression using a four-parameter sigmoid tanh function

$$y = \frac{1 + \tanh\left(\frac{2x-c}{w}\right)}{2\left(\frac{1}{y_{\max} - y_{\min}}\right)} + y_{\min} \quad (1)$$

where c is the centre of the cline, w is the cline width and $y_{\max}$ and $y_{\min}$ are the maximum and minimum trait values (that is, the trait values at the tails of the cline). This uses the cline model of Szymura and Barton (ref. 47), except that the minimum and maximum trait values are free to take on any value. Parameter searches were done using the solver function in Excel using a least-squares optimality criterion. Solver was run using the generalized reduced gradient (GRG) nonlinear algorithm with the following settings: convergence = 0.0001; central derivatives; multistart on; population size = 100.

To evaluate whether the data were adequately described by a 'flat' model (constant trait value across the transect) or a linear model (smooth transition), we fit these models, in addition to the sigmoid model, as candidate models. A flat model consists of a single parameter (population mean) defined as the grand mean of all individuals and is invariant across the transect. A linear model has two parameters, slope and y intercept, and was fit with linear regression. To evaluate which of the three models (one-parameter flat, two-parameter linear or four-parameter sigmoid) was a better fit to the data, we calculated ΔAICc and Akaike weights (w_i) for each model (methods following ref. 48) using the residual sum of squares divided by the sample size as the likelihood criterion.

Confidence intervals on parameter estimates were calculated using a Monte Carlo resampling method using the software GraphPad Prism. Briefly, this procedure involves the following steps: first, data are simulated for each observed x value using best-fit parameters of the observed cline, with scatter added by drawing data points randomly from a hypothetical normally distributed population with a s.d. equal to the observed $S_{y.x}$ (s.d. of the residuals). A cline is then fit to the simulated data and best-fit values of each parameter are recorded. This process is then repeated for a number of iterations, each time generating a new simulated data set, fitting a cline to that data set and recording parameter estimates. By using observed x values (that is, actual sampling locations) and observed residual variation, we are essentially simulating the distribution of cline parameter estimates that we might observe if we resampled the entire transect multiple times. Simulations were run for 10,000 generations and 95% confidence intervals were calculated on the simulation parameter estimates.

We tested for common centre (coincidence) and common width (concordance) among clines using global nonlinear regression. This method compares model fit when certain parameters are shared versus unshared among different variables. Under a scenario where all variables shift at the same position on the transect, a common centre parameter can be fit across all measured variables without a substantial reduction in model fit. This is expected, for example, in a scenario where there is a shift in the selective regime for one mimetic colour–pattern in one area versus another area along the transect. If the width of a cline on a phenotypic trait is a function of the strength of selection on that trait, a global width parameter may be expected when selection acts at similar strength on all traits; however, if selection is strong on some traits and weak on others, this will cause different cline widths and thus a common width parameter will not adequately fit all the data. A common width may also be expected when linkage disequilibrium is high in the centre of the hybrid zone, as is observed here (Supplementary Fig. 3). We evaluated four models representing different combinations of shared and unshared parameters (Supplementary Table 4): (a) no constraint (each variable with unique centre and width), (b) centres constrained to be equal, width unconstrained, (c) width constrained to be equal, centres unconstrained, and (d) centre constrained to be equal and width constrained to be equal. Best-fit shared parameter searches were done by fitting shared parameters to all data sets simultaneously, while unshared parameters were free to take on unique values for each data set. Goodness of fit was assessed by calculating ΔAICc for each model.

Mate-choice experimental design and analysis. To test for morph-based mating preferences, we conducted triad mate-choice experiments in which we introduced two females (one of each morph) into the terrarium of a given male for 1 h, and measured the amount of courtship time between the male and the varadero female versus the male and the striped female. For details on the populations we sampled, as well as details on husbandry and experimental protocols, see Supplementary Methods. In *R. imitator*, courtship is usually initiated when a calling male approaches a female. The female may then either reciprocate by following the male to a suitable oviposition site while the male continues calling, or show no interest[49]. The conditions of our mate-choice experiments allowed these behaviours to take place in that a male was free to initiate courtship with either female, and the female was free to reciprocate interest or not. Male initiation of courtship is readily observable in captivity as males (a) initiate a courtship call (shorter and more rapid than an advertisement call) and/or (b) begin to move in a staccato-like walk, often moving their rear legs erratically. Thus, when a male engaged in either of these behaviours in the vicinity of a female, this marked the initiation of courtship. Courtship was deemed to have ended under the following conditions: (a) the male, having initiated courtship, moves away and the female does not pursue or (b) the female moves away and the male does not pursue. A trial was excluded when one or both females remained hidden in the gravel during the trial, thus precluding any possibility for choice. Using these criteria, we measured in each trial the total amount of courtship between the male and the varadero female versus the male and the striped female.

Typically, in this kind of experimental setup, the two females to be introduced to the male would be matched for mass to control for any confounding effects of mass on preference. However, in this case, matching for mass was not feasible because the varadero population is larger than either striped population (striped-allopatric females $\bar{x} = 0.56$ g, s.d. $= 0.05$ g, $n = 43$; striped-transition females $\bar{x} = 0.56$ g, s.d. $= 0.04$ g, $n = 18$; varadero females $\bar{x} = 0.56$ g, s.d. $= 0.05$ g, $n = 30$), severely limiting the number of potential female combinations (for example, only the four heaviest striped-transition females would have qualified to be matched with the six smallest varadero females). To control for differences between females, we used a paired-samples' experimental design whereby a given pair of females was presented to a male of each morph. This design therefore addresses the question of how changing male morph type alters courtship probabilities when female identity is held constant.

To analyse mate-choice data, we used GLMM using the glmmADMB package[50] in R version 3.0.2 (ref. 51) with an underlying beta-binomial error distribution to test whether the time males spent courting each female morph was influenced by male population origin. 'Pair ID' (that is, a unique identifier assigned to each female pair) was used as a random effect to account for the paired-samples' experimental design. Following a significant result of the overall GLMM, we conducted *post hoc* tests to determine: (1) whether courtship preferences differ among morphs (specifically, comparing striped allopatric with varadero, and striped transition with varadero) and (2) whether courtship preferences differ among populations of the same morph (comparing striped allopatric to striped transition). *Post hoc* tests were run using the same GLMM procedure described above, except that we restricted the analysis to the populations of interest. To account for multiple comparisons, we adjusted *P* values using a FDR protocol[52] accounting for the fact that we conducted three *post hoc* tests. The protocols we used were approved by East Carolina University's Institutional Animal Care and Use Committee (AUP permit #D225a) before the start of this study.

Landscape genetics. We used a causal modelling framework[33,53] to test specific hypotheses of how geographic distance and colour–pattern differentiation between populations are associated with genetic distance. In landscape genetics, causal modelling uses Mantel tests and partial Mantel tests to evaluate alternative models explaining genetic distance between populations. Each model carries a set of statistical predictions; the model with all its predictions upheld is the one with the strongest support. In an IBD scenario, a significant correlation is expected between a geographic distance matrix (independent variable set) and a genetic distance matrix (dependent variable set). By using partial Mantel tests, the correlation between two dissimilarity matrices can be quantified while controlling for the effect of a third covariant matrix. For example, a partial Mantel test between colour–pattern distance and genetic distance with geographic distance as covariant matrix tests for the correlation between colour–pattern distance and genetic distance after the effects of geographic distance are removed. We used one measure of geographic distance (Euclidean distance), one measure of genetic distance (Nei's *D*) and one measure of colour–pattern distance (difference in discriminant score, see below) to test three models of genetic isolation. Details on each model and their associated predictions are given in Supplementary Table 3.

Causal modelling is often used to test how various landscape factors influence genetic isolation among populations[53,54]. This can be useful for species occupying heterogeneous habitats, where straight-line distance between populations may not be the most likely corridor of gene flow. However, in our case, all populations of *R. imitator* are from a contiguous lowland rainforest habitat without any geographic barriers separating populations. The only two substantial barriers in this area, the Huallaga River and the Cordillera Escalera Mountains, are located to the east and south, respectively, of all sampling sites. Therefore, for the geographic distance matrix, we simply calculated pairwise straight-line distance between populations. For the genetic distance matrix, we calculated Nei's genetic distance (D') between all pairs of populations in GenAlEx version 6.5 (ref. 55). To generate a colour–pattern distance matrix, we calculated pairwise differences in discriminant score from the kernel discriminant function analysis. Thus, because this analysis takes into account features of the model species, it can be thought of as a composite difference in mimetic colour–pattern. In addition to causal modelling, we used a multiple-matrix regression method[34] to quantify the relative effects of geographic distance and colour–pattern distance on genetic distance. This method is similar to Mantel and partial Mantel tests but incorporates multiple regressions, such that the relative effects of two or more predictor variables on genetic distance can be quantified, as can the overall fit of the model. Multiple-matrix regression was run with 10,000 permutations using the R script provided in ref. 34.

For details on microsatellite-genotyping protocols, Structure analyses and factorial correspondence analyses, see Supplementary Methods and Supplementary Table 6.

Bioacoustics. We recorded the advertisement calls of 58 *R. imitator* from eight localities. These localities are all located on the colour–pattern/microsatellite transect spanning the transition zone and thus can be used for cline analysis. Calls were recorded on a Marantz PMD660 solid state recorder using a Sennheiser ME 66-K6 microphone and analysed in Raven Pro version 1.3 (ref. 56). We quantified advertisement calls by measuring the following parameters: note length (measured from the start of the first pulse to the end of the last pulse), pulse rate (defined as pulse count divided by note time) and dominant frequency (the frequency at which peak amplitude is registered). For each male, a recording generally consisted of several notes. Measurements were always taken on at least three notes and then averaged for each male. As temperature is known to have a strong influence on certain aspects of amphibian calls[19], we took temperature measurements alongside each call recording in the same microhabitat as the calling male, which we then used to standardize call parameters by calculating regression residuals against temperature. We fit clines to each call parameter separately, plus a 'composite' call score that we calculated using a linear discriminant analysis.

References

1. Nosil, P. *Ecological Speciation* (Oxford Univ. Press, 2012).
2. Hatfield, T. & Schluter, D. Ecological speciation in sticklebacks: environment-dependent hybrid fitness. *Evolution* **53,** 866–873 (1999).
3. McKinnon, J. S. *et al.* Evidence for ecology's role in speciation. *Nature* **429,** 294–298 (2004).
4. Chamberlain, N. L., Hill, R. I., Kapan, D. D., Gilbert, L. E. & Kronforst, M. R. Polymorphic butterfly reveals the missing link in ecological speciation. *Science* **326,** 847–850 (2009).
5. Via, S. Natural selection in action during speciation. *Proc. Natl Acad. Sci. USA* **106,** 9939–9946 (2009).
6. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates Sunderland, 2004).
7. Alexandrou, M. A. *et al.* Competition and phylogeny determine community structure in Müllerian co-mimics. *Nature* **469,** 84–88 (2011).
8. Marek, P. E. & Bond, J. E. A Müllerian mimicry ring in Appalachian millipedes. *Proc. Natl Acad. Sci. USA* **106,** 9755–9760 (2009).
9. Greene, H. W. & McDiarmid, R. W. Coral snake mimicry: does it occur? *Science* **213,** 1207–1212 (1981).
10. Plowright, R. & Owen, R. E. The evolutionary significance of bumble bee color patterns: a mimetic interpretation. *Evolution* **34,** 622–637 (1980).
11. Symula, R., Schulte, R. & Summers, K. Molecular phylogenetic evidence for a mimetic radiation in Peruvian poison frogs supports a Müllerian mimicry hypothesis. *Proc. R. Soc. Lond. B Biol. Sci.* **268,** 2415–2421 (2001).

12. Turner, J. in: *Ecological genetics and evolution* 224–260 (Springer, 1971).
13. Bates, H. W. Contributions to an Insect Fauna of the Amazon Valley. Lepidoptera: Heliconidae. *Trans. Linn. Soc. Lond.* **23,** 495–566 (1862).
14. Kronforst, M. R. *et al.* Linkage of butterfly mate preference and wing color preference cue at the genomic location of *wingless*. *Proc. Natl Acad. Sci. USA* **103,** 6575–6580 (2006).
15. Jiggins, C., McMillan, W., King, P. & Mallet, J. The maintenance of species differences across a *Heliconius* hybrid zone. *Heredity* **79,** 495–505 (1997).
16. Arias, C. F. *et al.* A hybrid zone provides evidence for incipient ecological speciation in *Heliconius* butterflies. *Mol. Ecol.* **17,** 4699–4712 (2008).
17. Mallet, J. *et al.* Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics* **124,** 921–936 (1990).
18. Arias, C. F. *et al.* Sharp genetic discontinuity across a unimodal *Heliconius* hybrid zone. *Mol. Ecol.* **21,** 5778–5794 (2012).
19. Myers, C. W. & Daly, J. W. Preliminary evaluation of skin toxins and vocalizations in taxonomic and evolutionary studies of poison-dart frogs (Dendrobatidae). *Bull. Am. Mus. Nat. Hist.* **157,** 177–262 (1976).
20. Summers, K., Cronin, T. W. & Kennedy, T. Variation in spectral reflectance among populations of *Dendrobates pumilio*, the strawberry poison frog, in the Bocas del Toro Archipelago, Panama. *J. Biogeogr.* **30,** 35–53 (2003).
21. Brown, J. L. *et al.* A taxonomic revision of the Neotropical poison frog genus *Ranitomeya* (Amphibia: Dendrobatidae). *Zootaxa* **3083,** 1–120 (2011).
22. Silverstone, P. A. A revision of the poison-arrow frogs of the genus Dendrobates Wagler. *Nat. Hist. Mus. Los Angel. Cty. Sci. Bull.* **21,** 1–55 (1975).
23. Maan, M. E. & Cummings, M. E. Sexual dimorphism and directional sexual selection on aposematic signals in a poison frog. *Proc. Natl Acad. Sci. USA* **106,** 19072–19077 (2009).
24. Comeault, A. & Noonan, B. Spatial variation in the fitness of divergent aposematic phenotypes of the poison frog, *Dendrobates tinctorius*. *J. Evol. Biol.* **24,** 1374–1379 (2011).
25. Tazzyman, S. J. & Iwasa, Y. Sexual selection can increase the effect of random genetic drift—a quantitative genetic model of polymorphism in *Oophaga pumilio*, the Strawberry poison-dart frog. *Evolution* **64,** 1719–1728 (2010).
26. Hegna, R. H., Saporito, R. A. & Donnelly, M. A. Not all colors are equal: predation and color polytypism in the aposematic poison frog *Oophaga pumilio*. *Evol. Ecol.* **27,** 1–15 (2012).
27. Richards-Zawacki, C. L., Yeager, J. & Bart, H. P. No evidence for differential survival or predation between sympatric color morphs of an aposematic poison frog. *Evol. Ecol.* **27,** 783–795 (2013).
28. Yeager, J., Brown, J. L., Morales, V., Cummings, M. & Summers, K. Testing for selection on color and pattern in a mimetic radiation. *Curr. Zool.* **58,** 668–676 (2012).
29. Twomey, E. *et al.* Phenotypic and genetic divergence among poison frog populations in a mimetic radiation. *PLoS ONE* **8,** e55443 (2013).
30. Yeager, J. *Quantification of Resemblance in a Mimetic Radiation*. MS thesis, East Carolina University, Department of Biology (2009).
31. Barton, N. H. & Hewitt, G. Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* **16,** 113–148 (1985).
32. Summers, K., Cronin, T. & Kennedy, T. Cross-breeding of distinct color morphs of the strawberry poison frog (*Dendrobates pumilio*) from the Bocas del Toro Archipelago, Panama. *J. Herpetol.* **38,** 1–8 (2004).
33. Wang, I. J. & Summers, K. Genetic structure is correlated with phenotypic divergence rather than geographic isolation in the highly polymorphic strawberry poison-dart frog. *Mol. Ecol.* **19,** 447–458 (2010).
34. Wang, I. J. Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution* **67,** 3403–3411 (2013).
35. Gavrilets, S. *Fitness Landscapes and the Origin of Species* (Princeton Univ. Press, 2004).
36. Brown, J. L., Morales, V. & Summers, K. A key ecological trait drove the evolution of biparental care and monogamy in an amphibian. *Am. Nat.* **175,** 436–446 (2010).
37. Jiggins, C. D., Naisbit, R. E., Coe, R. L. & Mallet, J. Reproductive isolation caused by colour pattern mimicry. *Nature* **411,** 302–305 (2001).
38. Kronforst, M., Young, L. & Gilbert, L. Reinforcement of mate preference among hybridizing *Heliconius* butterflies. *J. Evol. Biol* **20,** 278–285 (2007).
39. McMillan, W. O., Jiggins, C. D. & Mallet, J. What initiates speciation in passion-vine butterflies? *Proc. Natl Acad. Sci. USA* **94,** 8628–8633 (1997).
40. Gomez, D. AVICOL, a program to analyse spectrometric data. Free executable available at http://sites.google.com/site/avicolprogram/or from the author at dodogomez@yahoo.fr (2006).
41. Endler, J. A. On the measurement and classification of colour in studies of animal colour patterns. *Biol. J. Linn. Soc.* **41,** 315–352 (1990).
42. Koenderink, J. J. & van Doorn, A. J. Surface shape and curvature scales. *Image Vis. Comput.* **10,** 557–564 (1992).
43. Dalal, N. & Triggs, B. in: *Computer Vision and Pattern Recognition, 2005. (CVPR 2005) IEEE Comput. Soc.* Vol. 1, 886–893 (IEEE, 2005).
44. Larsen, A. B., Vestergaard, J. S. & Larsen, R. HEp-2 cell classification using shape index histograms with donut-shaped spatial pooling. *IEEE Trans. Med. Imaging* **33,** 1573–1580 (2014).
45. Lindeberg, T. Scale-space: a framework for handling image structures at multiple scales. *CERN Eur. Organ. Nucl. Res. -Rep.* 27–38 (1996).
46. Mika, S., Rätsch, G., Weston, J., Schölkopf, B. & Müller, K. in: *Neural Networks for Signal Processing. IEEE Signal Process. Soc. Workshop* 41–48 (IEEE, 1999).
47. Szymura, J. M. & Barton, N. H. Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* **40,** 1141–1159 (1986).
48. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach* (Springer, 2002).
49. Brown, J. L., Twomey, E., Morales, V. & Summers, K. Phytotelm size in relation to parental care and mating strategies in two species of Peruvian poison frogs. *Behaviour* **145,** 1139–1165 (2008).
50. Skaug, H., Fournier, D., Nielsen, A., Magnusson, A. & Bolker, B. glmmADMB: generalized linear mixed models using AD Model Builder. *R Package Version 06* **5,** r143 (2011).
51. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Found. Stat. Comput. (2005).
52. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57,** 289–300 (1995).
53. Cushman, S. A., McKelvey, K. S., Hayden, J. & Schwartz, M. K. Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *Am. Nat.* **168,** 486–499 (2006).
54. Richards-Zawacki, C. L. Effects of slope and riparian habitat connectivity on gene flow in an endangered Panamanian frog, Atelopus varius. *Divers. Distrib.* **15,** 796–806 (2009).
55. Peakall, R. & Smouse, P. E. GENALEX 6: genetic analysis in excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6,** 288–295 (2006).
56. Charif, R., Waack, A. & Strickman, L. . *Raven Pro 1.3 user's manual* (Ithaca NY Cornell Lab of Ornithology, 2008).

Acknowledgements

We thank Jesse Delia and Jeff McKinnon for discussions and comments on the manuscript; Anders B.L. Larsen for discussion on image texture quantification; Morgan Kain for help with multiple-matrix regression; and Justin Touchon for help with the generalized linear mixed model analysis. We also thank Jason Brown, Santiago Cisneros, César Lopez, Manuel Guerra-Panaifo, Michael Mayer, Mark Pepper, Neil Rosser, Manuel Sanchez-Rodriguez, Lisa Schulte, Adam Stuckert, James Tumulty and Justin Yeager for help in the field. For help with research permits and museum specimens, we thank Pablo Venegas. This research was funded by a NSF DDIG (1210313) grant awarded to E.T. and K.S., a National Geographic Society grant awarded to K.S. (8751-10) and the NCCB scholarship (2012) at East Carolina University awarded to E.T. Research permits were obtained from the Ministry of Natural Resources (DGFFS) in Lima, Peru (Authorizations No. 050-2006-INRENA-IFFS-DCB, No. 067-2007-INRENA-IFFS-DCB, No. 005-2008-INRENA-IFFS-DCB). Tissue exports were authorized under Contrato de Acceso Marco a Recursos Geneticos No. 0009-2013-MINAGRI-DGFFS/DGEFFS, with CITES permit number 003302. All research was conducted following an animal use protocol approved by the Animal Care and Use Committee of East Carolina University.

Author contributions

E.T. and K.S. designed the study. E.T. performed the field sampling and experiments, and collected and analysed data. J.S.V. developed the computer-automated pattern extraction methods and analysed the data. E.T., J.S.V. and K.S. wrote the paper.

Additional information

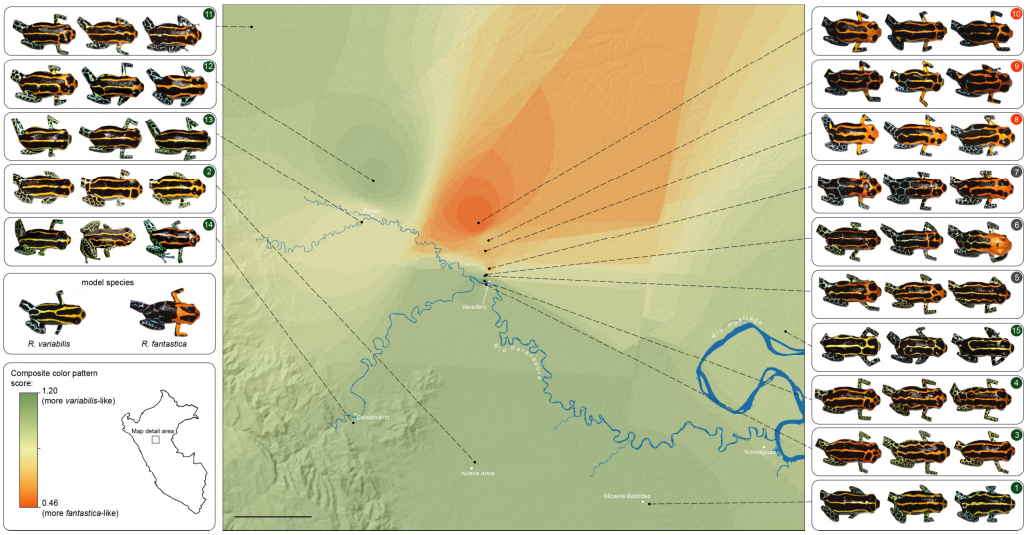
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

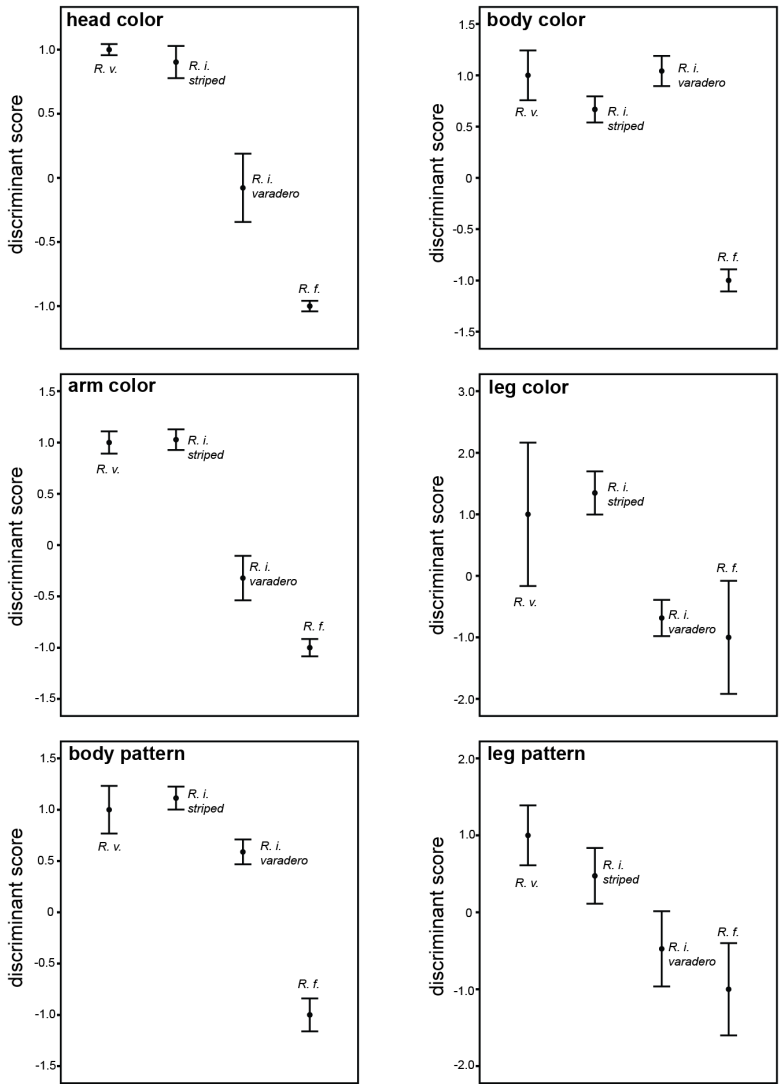
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions.

How to cite this article: Twomey, E. *et al.* Reproductive isolation related to mimetic divergence in the poison frog *Ranitomeya imitator*. *Nat. Commun.* 5:4749 doi: 10.1038/5749 (2014).

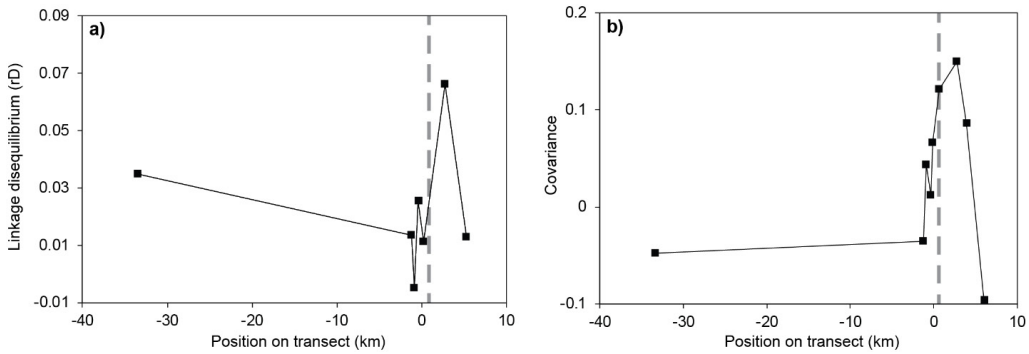
Supplementary Information



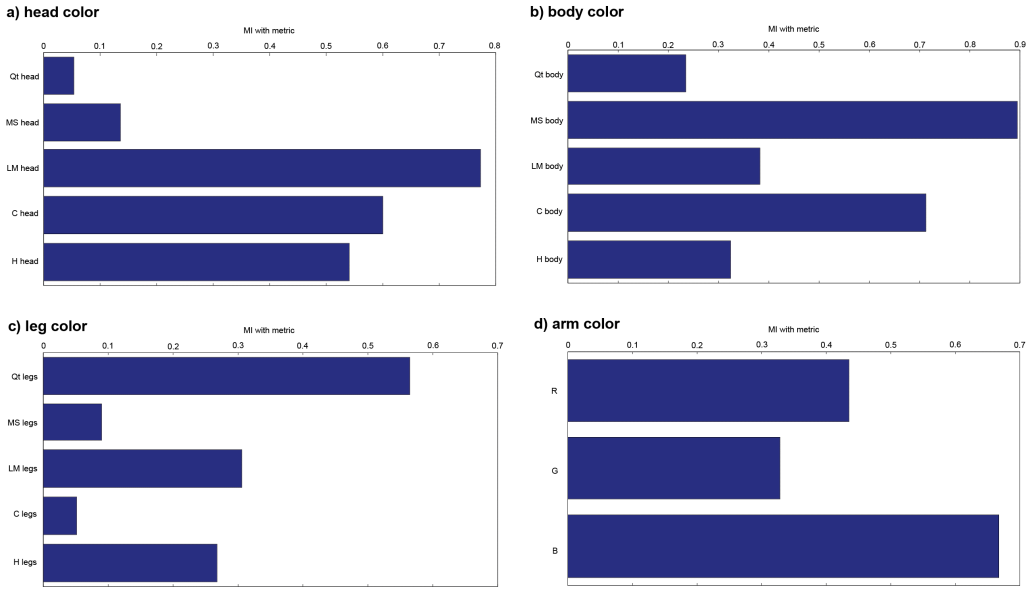
Supplementary Figure 1 – Sampling localities and variation in *Ranitomeya imitator*. Numbered boxes show *R. imitator* variation at a given locality (numbers correspond to localities in Supplementary Table 1). The colour of the dot indicates the putative morph: green dots indicate the striped morph (*R. variabilis* mimic), orange dots indicate the varadero morph (*R. fantastica* mimic) and grey dots indicate the transitional form between the two morphs. Model species are shown in middle-left box. Composite colour pattern score was calculated from the colour pattern data using a kernel discriminant function analysis, with the model species representing the training groups. Mean discriminant scores at each sampling locality were then interpolated for visualisation on the map using IDW interpolation in ArcGIS. Scale bar equals 10 km.



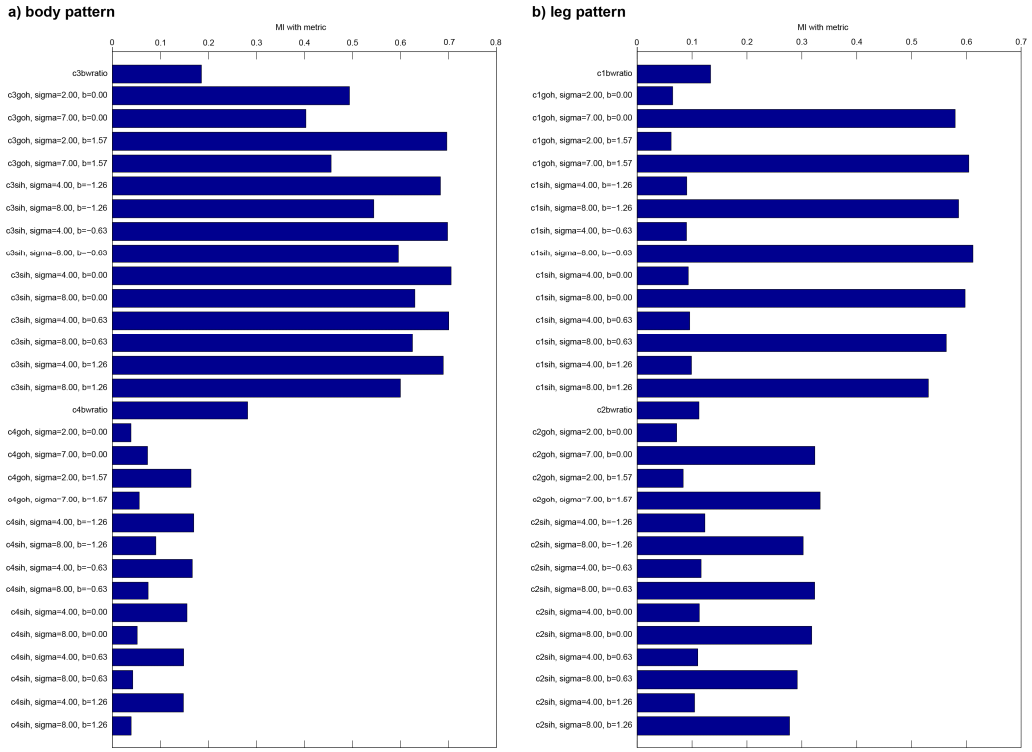
Supplementary Figure 2 – Mimicry comparison between model species and *R. imitator*. In each panel the average discriminant scores (±1 s.e.m.) from the kernel discriminant analysis are compared among model species (*R. variabilis*, denoted *R. v.*; *R. fantastica,* denoted *R. f.*) and two populations of *R. imitator* (*R. i. striped* and *R. i. varadero).* The two *R. imitator* populations plotted here are Micaela Bastidas and Varadero Forest 1 (see Supplementary Table 1), which are representative of "pure" striped and varadero morphs, respectively.



Supplementary Figure 3 – Linkage disequilibrium and phenotypic covariance across the sampling transect. (**a**) Multilocus linkage disequilibrium ($\bar{r}_d$) and (**b**) covariance between arm colour and leg colour in each population was calculated plotted along the transect (x-axis). The dashed grey lines show the position of the colour pattern cline centre. For both linkage disequilibrium and phenotypic covariance, the peaks occur near the centre of the colour pattern cline, consistent with the predictions of a hybrid zone.



Supplementary Figure 4 – Mutual information plot for colour variables. See methods for details on how mutual information is calculated. In general, larger bars indicate greater variable contribution to the discriminant function. For panels (**a–c**), variables were derived from the Avicol analyses on the spectrometer data. Variable prefixes are as follows: Qt = brightness, MS = blue/yellow axis position, LM = red/green axis position, C = chrominance, H = hue. For panel (**d**), variables were derived from measuring the intensities of the red (R), green (G), and blue (B) channels in Photoshop.



Supplementary Figure 5 – Mutual information plot for pattern variables. Mutual information calculations are detailed in the methods. As in Supplementary Figure 4, larger bars indicate greater variable contribution to the discriminant function. Panels are split by body region: (**a**) body pattern and (**b**) leg pattern. Variables use the following naming convention: body region, metric, and associated parameters. For example, c3goh, sigma=2, b=0 indicates that the body region of interest was c3, the metric was goh (gradient orientation histogram), and the associated parameters for its extraction were sigma=2 and b=0. Body regions are as follows: c3=lower dorsum, c4=head, c1=right leg, c2=left leg. Extracted metrics are as follows: goh=gradient orientation histogram, sih=shape index histogram, bwratio=colour/non-colour ratio.

Supplementary Table 1 – Sampling localities and sample sizes for each kind of data collected for the study. Numbers in the first column correspond to localities shown on Supplementary Fig. 1.

locality	number on Suppl. Fig. 1	position on transect (km from centre)	latitude	longitude	colour pattern *n*	microsatellites *n*	male mass *n*	advertisement call *n*
Micaela Bastidas	1	-33.46	-5.9554	-76.2424	31	36	23	11
Nuevo Arica	2	-22.63	-5.9107	-76.4280	6	—	—	—
Varadero - South Bank	3	-1.25	-5.7177	-76.4163	8	22	13	13
Varadero - Bridge	4	-0.92	-5.7142	-76.4178	6	6	4	4
Varadero - Stream	5	-0.39	-5.7084	-76.4174	16	16	11	6
Varadero - Transition 1	6	-0.16	-5.7073	-76.4161	5	5	3	4
Varadero - Transition 2	7	0.64	-5.7009	-76.4128	4	4	2	—
Varadero - Forest 1	8	2.72	-5.6821	-76.4171	25	36	22	17
Varadero - Forest 2	9	3.92	-5.6710	-76.4137	3	3	—	—
Varadero - Forest 3	10	6.03	-5.6515	-76.4241	5	5	—	3
Monte Cristo	11	not included in transect	-5.4395	-76.6655	6	—	—	—
Panan North	12	not included in transect	-5.6067	-76.5361	3	—	—	—
Panan South	13	not included in transect	-5.6510	-76.5484	4	—	—	—
Balsapuerto	14	not included in transect	-5.8542	-76.5431	—	3	—	—
Bajo Huallaga	15	not included in transect	-5.7618	-76.0780	5	—	—	—

Supplementary Table 2 – Model fit results for each of three candidate models describing transect variation in all six variables measured along the transect. In all cases, the best supported model (as indicated by AICc) is shown in bold. For cases where the sigmoid model was best supported, point estimates and 95% Monte Carlo confidence intervals on centre and width parameters are given.

variable	model	AICc	ΔAICc	Akaike weight	centre	centre 95% CI	width	width 95% CI
arm colour	flat	-89.5	134.0	0.000	—	—	—	—
	linear	-150.2	73.3	0.000	—	—	—	—
	sigmoid	**-223.4**	**0.0**	**1.000**	**0.72**	**0.02 – 1.65**	**4.15**	**1.79 – 6.43**
body colour	flat	-181.8	15.9	0.000	—	—	—	—
	linear	**-197.7**	**0.0**	**0.969**	—	—	—	—
	sigmoid	-190.8	6.9	0.031	—	—	—	—
head colour	flat	-55.8	58.6	0.000	—	—	—	—
	linear	**-114.4**	**0.0**	**0.894**	—	—	—	—
	sigmoid	-110.1	4.3	0.106	—	—	—	—
leg colour	flat	46.9	85.2	0.000	—	—	—	—
	linear	20.5	58.8	0.000	—	—	—	—
	sigmoid	**-38.3**	**0.0**	**1.000**	**0.77**	**0.04 – 1.93**	**2.38**	**0.03 – 4.86**
body pattern	flat	-191.6	44.0	0.000	—	—	—	—
	linear	-203.3	32.4	0.000	—	—	—	—
	sigmoid	**-235.6**	**0.0**	**1.000**	**0.03**	**-0.17 – 1.82**	**0.10**	**0.00 – 4.75**
leg pattern	flat	13.0	2.9	0.180	—	—	—	—
	linear	**10.1**	**0.0**	**0.772**	—	—	—	—
	sigmoid	15.6	5.6	0.048	—	—	—	—
microsatellites FCA axis 1	flat	-138.8	145.6	0.000	—	—	—	—
	linear	-164.8	119.6	0.000	—	—	—	—
	sigmoid	**-284.4**	**0.0**	**1.000**	**0.31**	**-0.15 – 0.63**	**0.64**	**0.01 – 1.54**
male mass	flat	-423.3	82.0	0.000	—	—	—	—
	linear	-424.5	80.9	0.000	—	—	—	—
	sigmoid	**-505.3**	**0.0**	**1.000**	**-0.13**	**-0.15 – 0.63**	**0.07**	**0.00 – 1.50**
advertisement call	flat	114.8	39.2	0.000	—	—	—	—
	linear	102.5	26.9	0.000	—	—	—	—
	sigmoid	**75.6**	**0.0**	**1.000**	**-0.19**	**-0.5 – 1.91**	**0.40**	**0.00 – 5.41**

Supplementary Table 3 – Causal modelling results of factors potentially influencing genetic distance and associated statistical predictions under each hypothesis. Results from the partial Mantel tests are given as a p-value and a yes/no indication of whether the prediction was supported. For each statistical prediction, a × separates the two dependent matrices, with the covariate matrix separated by a period. For example, Dist × Gen . Cp tests for the correlation between Dist (geographic distance) and Gen (genetic distance) controlling for the effect of Cp (colour pattern distance).

Factor(s) influencing genetic structure	Statistical predictions	Result (p-value)	Prediction supported?
Geographic distance	Dist × Gen . Cp = sig.	0.434	no
	Cp × Gen . Dist = n.s.	0.011	no
Colour pattern	Cp × Gen . Dist = sig.	0.011	yes
	Dist × Gen . Cp = n.s.	0.111	yes
Geographic distance and colour pattern	Cp × Gen . Dist = sig.	0.010	yes
	Dist × Gen . Cp = sig.	0.115	no

Supplementary Table 4 – Global regression results for common centre and common width parameters for the six variables showing sigmoidal variation across the transect. Models were compared globally using AICc. The best supported model (shared centre) is shown in bold.

		a) No constraints			**b) Shared centre**			c) Shared width			d) Shared centre and width		
	n	centre	width	RSS	**centre**	**width**	**RSS**	centre	width	RSS	centre	width	RSS
arm colour	108	0.72	-4.14	29.18	**0.52**	**-3.68**	**29.24**	0.00	-0.35	32.55	0.22	-0.97	31.09
leg colour	108	0.77	-2.38	45.85	**0.52**	**-1.86**	**45.94**	0.66	-0.35	46.84	0.22	-0.97	47.26
body pattern	108	0.03	-0.10	67.10	**0.52**	**-0.10**	**67.11**	0.21	-0.35	67.12	0.22	-0.97	67.56
male mass	78	-0.12	-0.07	24.74	**0.52**	**-0.10**	**24.89**	0.07	-0.35	24.83	0.22	-0.97	26.46
call	58	-0.22	-0.31	26.20	**0.52**	**-4.15**	**27.28**	-0.22	-0.35	26.21	0.22	-0.97	28.60
microsatellites	133	0.31	-0.64	14.72	**0.52**	**-0.27**	**14.76**	0.48	-0.35	14.76	0.22	-0.97	14.94
total RSS				207.79			**209.23**			212.31			215.92
parameters				24			**19**			19			14
AICc				-571.74			**-578.44**			-569.78			-570.38
ΔAICc				6.70			**0.00**			8.66			8.06

Supplementary Table 5 – Parameters used in the kernel discriminant analysis.

	Regularization λ	*Kernel width* γ
head color	0.002	2.14
body color	0.001	9.00
leg color	0.310	6.30
arm color	0.000	4.78
leg pattern	0.010	10.00
body pattern	0.001	22.90

Supplementary Table 6 – Primer sequences for microsatellites.

locus	Forward primer sequence	Reverse primer sequence
RimiA06	CTTAATTGAGTAATTGTCAAG	GCTTTTGGATAATCAGTATCG
RimiA07	TTCTTAATTGAGTAATTGTC	TCCTTAATATACCAGTTAAGC
RimiB01	TAATTGTATTTGTCACTGAC	ATTTTTGCGGGCATATTCGG
RimiB02a	TCGAGATTTTAGCAGTGTTTTATCC	CATGAAAACCATATTTCGGACA
RimiB07a	CACCGTGCACTGGTTATCTATC	GTTTCGCTCAACCCTAGTGC
RimiB11	GTAAGTCCGTATATGTCGATG	CCTGAGAGTGTAATGGATAGAC
RimiC05a	CGTTTCGCTCAACCCTAGTC	ATGGAGGCAATCCACAAATC
RvarD01	GAAAAAGCATTACAGCTCATCAA	GCCGAAACATTGCCATAAAT
RimiD04	CTCCAAAACACACCCCAAAC	AGAGGTGCTGCCCTTTTGTA
RimiE02a	GCAGAGGGGATTAGGGACTC	TGGGTAGCTGTGTTCCATGA
RimiF06	TTGATATTCTGAGGTATG	GTAGCTTATGGCAGCTACG

Supplementary Methods

Kernel discriminant analysis supplement. We regularized the kernel discriminant analysis solution with λ times the identity matrix and used a Gaussian kernel with width γ (Supplementary Table 5). These parameters were selected to minimize intra-location variance of the *R. imitator* discriminant scores while keeping them within the span of the two model species'

discriminant scores. The formulation of kernel discriminant analysis unfortunately makes it impossible to inspect, e.g., the loadings as one would do in linear discriminant analysis to determine variable importance. However, the mutual information (MI) between each of the original variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ and the discriminant scores (metric) $\mathbf{z}$ can be inspected to elucidate which variables are relevant for discrimination between the model species. Mutual information is defined as

$$MI(x, y) = \frac{H(x) + H(y) - H(x, y)}{H(x) + H(y)} \tag{1}$$

where $H(x)$ and $H(x,y)$ are marginal and joint entropies respectively. Parzen window estimates of MI between the variables used to define each of the six metrics, and the resulting metric can be seen in Supplementary Figures 4 and 5. The values $MI(\mathbf{x}_i, \mathbf{z})$ are normalized such that $MI(\mathbf{x}_i, \mathbf{x}_i) = MI(\mathbf{z}, \mathbf{z}) = 1$. For colour data (Supplementary Fig. 4), we see that for head colour, red-green axis (LM), chroma (C), and hue (H) have relatively high MI, whereas for body colour blue-yellow axis (MS) and chroma have high MI. For leg colour, brightness (Qt) had the highest MI. For arm colour, blue channel intensity (B) had the highest MI. For pattern data (Supplementary Fig. 5), we see that for dorsal pattern, most variables associated with pattern variation on the lower dorsum (prefix c3) show high MI, whereas variables associated with pattern variation on the head (prefix c4) show lower MI. For leg pattern, both shape index histograms (sih) and gradient orientation histograms (goh) showed high MI on each leg (c1 and c2), although this depended mainly on the associated extraction parameters.

Mate choice: Animal collection, husbandry, and protocols. We used *Ranitomeya imitator* from three populations for mate choice experiments (GPS points given in Supplementary Table 1):

(1) *Striped-allopatric (site 1; Supplementary Fig. 1)* – Frogs from this population are mimics of *R. variabilis*, with yellow pinstripes along the dorsum and a pale greenish or bluish reticulum on the legs and venter. These frogs were collected from near Micaela Bastidas, a village 33 km to the southeast of the transition zone and represent the 'striped-allopatric' population in our analyses.

(2) *Striped-transition (sites 3, 4, and 5; Supplementary Fig. 1)* – These frogs are also *R. variabilis* mimics, with yellowish-orange stripes on the dorsum and pale greenish legs. Due to the difficulty of collecting striped frogs in this area because of deforestation, we collected frogs from

three different sites. Two of these sites (Varadero Stream and Varadero Bridge on Supplementary Table 1) are on the north side of the Paranapura river and one site (Varadero South Bank) is on the south side. We treated these three collecting sites as a single population for mate choice trials because the Paranapura river is small (50-80 m across in most places) and appears to present no noticeable barrier to gene flow. For example, the genetic distance (Nei's D') between Varadero South Bank and Varadero Stream (same morph, opposite sides of river) is 0.226 over 1 km, whereas the genetic distance between Varadero Stream and Varadero Forest 1 (different morphs, same side of river) is 0.572 over 2.9 km. For comparison, this is roughly equivalent to the genetic distance between Varadero Stream and Micaela Bastidas ($D' = 0.580$), two striped populations separated by 33.4 km airline distance.

(3) *Varadero (site 8; Supplementary Fig. 1)* – This is the *R. fantastica* mimic morph, with orange dorsal colouration, orange upper arms, and navy blue reticulation on the legs and lower body. Frogs were collected from a single site 3.5 km north of the village of San Gabriel de Varadero. For consistency with previous studies[1,2] we refer to this morph as the varadero morph, despite the fact that striped frogs can also be found near this village.

Male and female *Ranitomeya imitator* were collected in the field and kept in captivity in Tarapoto, Peru. Sexual dimorphism in this species is subtle and mainly related to size[3], so when possible frogs were sexed based on behavioural observations made while collecting (e.g., calling, territorial fighting, tadpole transport, courtship behaviour). Frogs were weighed to the nearest 0.01 g, which was also useful for sex identification as females are generally heavier than males (across all known *R. imitator* populations, females: $\bar{x} = 0.60$ g, s.d. = 0.07 g, $n = 130$; males: $\bar{x} = 0.48$ g, s.d. = 0.06 g, $n = 176$). Frogs were housed individually in glass terraria (dimensions in cm 50 x 30 x 30). Terraria had roughly two inches of washed gravel as a substrate (primarily for temperature stability throughout the day), leaf litter, and were planted with two bromeliads (pineapple tops). Water and food (wild fruit flies) were both constantly available. For each of the three populations, we targeted a sample size of 20 responses to analyse mating preferences, as specified by the animal use protocol permit (AUP #225a).

Mate choice trials were initiated by releasing two females simultaneously into the terrarium of a male. Trials were filmed for one hour. After the trial, the same females were then released into a terrarium of a male of the opposite morph of the first male, and again filmed for one hour. To account for any order effects, the morph of the male tested first was determined at random. In cases where the male or one or more females were unresponsive (typically by hiding in the

gravel), trials were re-run at a later date. Terraria were illuminated with a full-spectrum ZooMed AvianSun 5.0 UVB 26 watt compact fluorescent bulb. To allow the full spectrum of light to pass into the terrarium, we constructed a special terrarium cover out of UV-transparent acrylic that we used during trials.

Population genetics. Genetic divergence between populations was assessed by genotyping 136 *R. imitator* individuals at 11 microsatellite loci. We amplified the following loci: RimiA06, RimiA07, RimiB01, RimiB02a, RimiB07a, RimiB11, RimiC05a, RvarD01, RimiD04, RimiE02a, and RimiF06 (see Supplementary Table 6 for primer sequences) following extraction and amplification protocols described in ref. 4, with the exception that 56°C was used as the annealing temperature for B07, C05a, and E02a and 54°C for D01. Forward primers were labelled with a fluorescent tag for visualisation (6-FAM, NED, PET, or VIC). Loci were amplified individually and multiplexed for sequencing. Sequencing was done on an ABI 3130 sequencer and fragment sizes were analysed using GeneMapper software (Applied Biosystems). We used Micro-Checker software version 2.2.3 (ref. 5) to check for presence of null alleles. Three out of the original eleven loci (A07, B07, E02) showed evidence for high null allele frequencies (mean across populations > 0.09), thus these loci were omitted from further analyses.

We used the program Structure version 2.3.4 (ref. 6) to investigate population genetic structure from the microsatellite data. This program employs a Bayesian clustering algorithm to assign individuals probabilistically to each of K populations, where K, the number of populations, is unknown. The program was run with a burn-in of 50,000 generations and 500,000 subsequent generations, from one to five genetic clusters (K = 1–5), with five replicates at each value of K. The program was run using the admixture model with allele frequencies correlated. No prior information on sampling location was used in the model. To determine the number of clusters that best describe the data, we used the method described in ref. 7, which is based on the second-order rate of change of the log-likelihood. This method was implemented using Structure Harvester[8].

To estimate cline shape for the microsatellite data, we used the first major axis from a factorial correspondence analysis (FCA), calculated using the software Genetix version 4.5 (ref. 9). This method is conceptually similar to principal components analysis, except it takes into account features of genetic data such as heterozygosity and homozygosity. This analysis was run

using the eight microsatellite loci and the nine localities on the sampling transect for which we had genetic data. The program was run without any population information on the samples.

Dispersal estimates. As individuals disperse into the centre of a hybrid zone, they carry with them the gene combinations characteristic of their parental populations[10]. This influx of parental genotypes into the hybrid zone creates linkage disequilibria among unlinked genetic loci, which are broken down through recombination in hybrids. Thus, strong linkage disequilibrium in a hybrid zone is evidence of reduced hybridization and increased reproductive isolation among parental types. Similarly, the influx of parental phenotypes into a hybrid zone will create covariance among independent phenotypic traits[11]. As with linkage disequilibrium, this phenotypic covariance should peak in the centre of the hybrid zone, and will be maximized when reproductive isolation is complete[11].

We used estimates of cline width, linkage disequilibrium, and phenotypic covariance to calculate the scale of dispersal (σ) in two different ways. First, we used the relationship (from ref. 11)

$$\sigma = \sqrt{\frac{r\overline{D}w^2}{1+r}} \tag{2}$$

where r is the recombination rate among loci (assumed to be 0.5 for unlinked loci), w is the width of the cline, and $\overline{D}$ is the peak linkage disequilibrium among genetic loci in the centre of the hybrid zone. For cline width, we used a value of 0.97 km, which corresponds to the point estimate for cline width for the entire dataset (see Supplementary Table 4, model D). To calculate linkage disequilibrium, we used the software Multilocus version 1.3 (ref. 12) to calculate the multilocus linkage disequilibrium estimator $\bar{r}_d$ which has a form similar to a correlation coefficient and can have a maximum value of 1. We found that, consistent with the predictions of a hybrid zone, linkage disequilibrium among the eight microsatellite loci peaked ($\bar{r}_d$ = 0.066) near the hybrid zone centre (Supplementary Fig. 3). Second, we used phenotypic covariance to estimate dispersal using the following relationship[11,13]

$$\sigma = \sqrt{\frac{2rC_{max}}{z1'_{max}z2'_{max}(1+r)}}$$

(3)

where $z1'_{max}$ and $z2'_{max}$ are the maximal slopes (defined as 4/width) of the clines on traits z1 and z2, respectively, C_{max} is the maximum covariance between traits z1 and z2, and r is the recombination rate (again assumed to be 0.5 for unlinked traits). We chose two phenotypic traits showing clear sigmoidal variation across the transect: arm colour and leg colour. These two traits have a maximum slope of 0.96 and 1.68, respectively, in the centre of the hybrid zone, and have a maximum covariance of 0.15, which occurs near the hybrid zone centre (Supplementary Fig. 3). Using linkage disequilibrium, we estimated the dispersal rate (σ) to be 0.095 km per generation. Using phenotypic covariance, we estimated $\sigma = 0.248$ km per generation. Although we have no direct survey data in *R. imitator* with which to compare these dispersal estimates, given the available field observations, these estimates seem reasonable. While adult *R. imitator* are highly territorial and occupy small home ranges (approximately 5–14 m^2, ref. 14), these home ranges are centred around reproductive resources and tightly packed in space such that most reproductive resources in a given site will be monopolized by a breeding pair. Thus, a juvenile or an individual in search of a territory may be forced to disperse a large distance in search of suitable breeding habitat. For example, two transient adult *R. imitator* moved distances of 19 and 23 m during a field study[14]. Additionally, one juvenile moved across an entire study plot (a distance of approximately 40 m) over the course of about one week (J. Brown, pers. comm.). Finally, during the course of a separate study[15], a male who was removed from his territory and released 160 m away was able to return to his territory in only five days (J. Tumulty, pers. comm.).

If the current hybrid zone reflects secondary contact with neutral diffusion, the width of the cline (w) can be predicted using the number of generations since contact (T) and the dispersal rate (σ), with the following equation[16]

$$w = \sqrt{(2\pi)}\,\sigma\,\sqrt{T} \tag{4}$$

Using our lower estimate of dispersal (0.095 km per generation), a cline formed by secondary contact with neutral diffusion should exceed the overall observed cline width (0.97 km; see Supplementary Table 4 model D) in only 17 generations. Using the upper estimate of dispersal (0.248 km per generation), a neutral cline would only take three generations to exceed the observed cline width. *Ranitomeya imitator* has a generation time of roughly eight months, which means secondary contact (assuming neutral diffusion) would have occurred not more than 11.3

years ago. Therefore, if the current hybrid zone is due to secondary contact, it is likely maintained by some barrier to gene flow (possibly assortative mating, divergent selection, or selection against hybrids), otherwise the cline is too narrow to be reasonably described by a neutral diffusion model.

Supplementary References

1. Yeager, J., Brown, J. L., Morales, V., Cummings, M. & Summers, K. Testing for selection on color and pattern in a mimetic radiation. *Curr. Zool.* **58,** (2012).
2. Twomey, E. *et al.* Phenotypic and genetic divergence among poison frog populations in a mimetic radiation. *PloS One* **8,** e55443 (2013).
3. Brown, J. L., Twomey, E., Morales, V. & Summers, K. Phytotelm size in relation to parental care and mating strategies in two species of Peruvian poison frogs. *Behaviour* **145,** 1139–1165 (2008).
4. Brown, J. L., Chouteau, M., Glenn, T. & Summers, K. The development and analysis of twenty-one microsatellite loci for three species of Amazonian poison frogs. *Conserv. Genet. Resour.* **1,** 149–151 (2009).
5. Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. & Shipley, P. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4,** 535–538 (2004).
6. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
7. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14,** 2611–2620 (2005).
8. Earl, D. & vonHoldt, B. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4,** 359–361 (2012).
9. Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. & Bonhomme, F. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. *Lab. Génome Popul. Interact. CNRS UMR* **5000,** (1996).

10. Szymura, J. M. & Barton, N. H. Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* **40,** 1141–1159 (1986).
11. Barton, N. H. & Gale, K. S. in *Hybrid Zones Evol. Process* 13–45 (1993).
12. Agapow, P.-M. & Burt, A. Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1,** 101–102 (2001).
13. Gay, L., Crochet, P.-A., Bell, D. A. & Lenormand, T. Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution* **62,** 2789–2806 (2008).
14. Brown, J. L., Morales, V. & Summers, K. Home range size and location in relation to reproductive resources in poison frogs (Dendrobatidae): a Monte Carlo approach using GIS data. *Anim. Behav.* **77,** 547–554 (2009).
15. Tumulty, J., Morales, V. & Summers, K. The biparental care hypothesis for the evolution of monogamy: experimental evidence in an amphibian. *Behav. Ecol.* **25,** 262–270 (2014).
16. Endler, J. A. *Geographic variation, speciation, and clines*. (Princeton University Press, 1977).

Paper D

Identifying pleiotropic control of adaptive phenotypes

Identifying pleiotropic control of adaptive phenotypes

Jacob S. Vestergaard[a], Evan Twomey[b], Allan A. Nielsen[a], Kyle Summers[b], Rasmus Nielsen[c]

[a]*Technical University of Denmark, Department of Applied Mathematics and Computer Science*
[b]*East Carolina University, Department of Biology*
[c]*University of California, Berkeley, Department of Integrative Biology*

Abstract

The genetic basis of complex traits that are polymorphic in an admixture zone has been the subject of considerable research, but is often difficult to characterize in species that are not amenable to controlled laboratory crosses. Here we develop a likelihood based approach to determine if two polymorphic phenotypes in an admixture zone are controlled by the same or different sets of genes. After evaluating the method using extensive simulations, we apply it to complex color pattern variation in the aposematic and mimetic frog *Ranitomeya imitator*. We show that patterns of banding and stripes, body color, and patterning on the leg are most likely controlled by a single set of genes. We also show that this can be efficiently produced using a simple Reaction-Diffusion model of pattern formation.

1. Introduction

The genetic basis of adaptive traits has been a topic of intense research focus over the past decade (e.g., Counterman *et al.*, 2010; Joron *et al.*, 2011; Kunte *et al.*, 2014; Martin *et al.*, 2012; Nachman *et al.*, 2003; Papa *et al.*, 2008), although the exact genetic basis has only been determined for a small number of traits (Martin and Orgogozo, 2013). Much attention has been focused on traits associated with aposematism and/or mimicry, in particular color pattern in the celebrated *Heliconius* system. In *Heliconius* butterflies color patterns on the wings warn potential predators that the butterflies are unpalatable. In the highly polymorphic *H. erato* and *H. melpomene* there are three components of the color pattern: the color of the forewing band (yellow or red), the presence or absence of a red patch on the proximal portion of the forewing, and the presence or absence of red hindwing rays (Sheppard *et al.*, 1985; Papa *et al.*, 2008). Although these elements are often reduced to two major phenotypes ('postman' and 'rayed'), there are in fact many different color forms and dozens of proposed subspecies of both *H. erato* and *H. melpomene*. Substantial progress has been made on understanding the genetic basis of the system. The red color variation is controlled by genes in a 400 KB region (Counterman *et al.*, 2010; Baxter *et al.*, 2010), with most of the variation due to expression differences in a gene, *optix*, located within the region. In addition, *WntA* contributes to pattern formation in forewing band shape (Martin *et al.*, 2012). Identifying these loci is the culmination of years of research involving a number of research groups. The first step in such a research program is to identify and quantify phenotypes and to establish the genetic basis of these traits, in terms of number of loci and covariance between different traits and different hypothesized loci. Most systems are still at a stage where this is a major challenge, in particular in organisms that are not amenable to laboratory crosses. One such species is the dendrobatid frog *Ranitomeya imitator*.

R. imitator forms a mimetic complex in northern Peru, where different populations have apparently evolved to resemble distinct model species in different regions (Symula *et al.*, 2001, 2003; Yeager *et al.*, 2012; Twomey *et al.*, 2013). In one case, *R. imitator* and its putative model (or co-mimic), *R. variabilis*, appear to undergo geographic change in color pattern in parallel, and it is unclear which species adverged on which (Chouteau *et al.*, 2011), but generally speaking phylogenetic analysis indicates that divergence in the putative model species is ancient relative to more recent divergence in *R. imitator*, indicating that this species has undergone a mimetic radiation and adverged in color pattern on distinct model species in different areas (Symula *et al.*, 2001, 2003; Yeager *et al.*, 2012; Twomey *et al.*, 2013). Across the range of *R. imitator*, a number of transition zones harbor intermediate forms that indicate the interbreeding of distinct morphs (Figure 1) and the production of hybrid intermediate forms (Twomey *et al.*, 2013). These transition zones are quite broad (e.g. 7 km) in some cases, indicating that interbreeding has been going on for long periods and is likely to contain multiple generations of color pattern intermediates. This provides a valuable opportunity for an analysis of the genetic basis of the control of color pattern, as the multiple intermediate forms produced across multiple generations of interbreeding provide the variation necessary to infer patterns of control using statistical methods of inference based on expected patterns

of co-inheritance for traits that are controlled by the same or different sets of genes. *Ranitomeya imitator* shows various combinations of pattern on the dorsal and ventral regions, and the extremities (i.e. fore and hind-legs). These include banding, striping, spotting, reticulation, as well as different combinations of color (brightness, hue and saturation). Dorsal coloration ranges from green, to yellow, to orange. Leg and arm coloration, in most populations, ranges from navy blue to greenish-blue. However, the banded population has distinctly orange legs. Dorsal pattern takes on three principal forms: parallel longitudinal stripes (striped morph), parallel latitudinal stripes (banded morph), and reticulated (spotted morph). Leg pattern in all populations, save one, is reticulated. The exception is the banded morph, which possesses no such leg reticulation, but rather thin stripes running lengthwise on the leg that appear to be extensions of the dorsal pattern. There are other subtle variations, for example, in the varadero morph, the navy blue reticulation on the legs extends up onto the lower dorsum. This morph also is unique in that it has less extensive melanization on the head, giving the appearance of a more "colorful" head, and this coloration extends down onto the upper arms, giving the appearance of an orange arm patch. While some covariation between these elements seems likely, a statistical approach is necessary to confirm or reject this hypothesis.

Figure 1: Four different morphs of *Ranitomeya imitator* and the respective model species with which they engage in Müllerian mimicry. *R. imitator* is on the left and the model species on the right.

We have previously devised methods for automatic quantification of phenotypes in this system have shown that the major mimetic phenotype is controlled by only one or a few genes (Vestergaard *et al.*, 2014). However, the *R. imitator* phenotype is highly complex with many different components. A major question in this, and many other systems, is the degree to which these different components are controlled by the same or different genes. This question can be addressed by analyses of genetic crosses between different morphological forms. Unfortunately, such crosses are not easily obtained in many organisms, including *R. imitator*. However, segregating variation in hybrid zones (also called introgression or admixture zones depending on context) can be used in lieu of controlled laboratory crosses. The major objective of this paper is to develop a new likelihood method for using individuals sampled from a hybrid zone to test if two or more quantitative traits are controlled by the same or different loci. After testing the new method and showing that it is statistically valid, we will then apply it to data from *R. imitator*, and use it to show that the major phenotypes, except leg color, are all controlled by the same gene(s).

This then raises the question as to how a single gene can control all these different phenotypes. One obvious explanation is a transcription factor, or other regulatory variant, that controls the activity of multiple downstream genes. Another possibility is a 'supergene' (Joron and Mallet, 1998; Kunte *et al.*, 2014), a tightly linked group of genes, perhaps fixed by an inversion, that together control the traits. A last possibility is that the phenotypes are pleitropic, or mechanistically co-inherited, because they depend on the same events during development. We will use a simple mathematical model of pattern formation in *R. imitator* to show that this latter explanation is a viable possibility.

The model we will use is a reaction-diffusion model. Reaction-diffusion models can be used to model the diffusion of proteins that control pattern formation during morphogenesis. They have previously been used extensively to model patterns of skin pigmentation in animals (Meinhardt, 1982; Murray, 2002). These mechanistic models can generate many possible patterns and have often been found to generate patterns similar to those ob-

served in nature. Obviously, the similarity between the patterns generated by these models and those observed in nature should not be interpreted as evidence that the mechanism generating the patterns in nature are the same as those implemented in the mathematical models. Nonetheless, there are examples of mechanisms underlying pattern formation that are adequately described by reaction-diffusion models, in particular pigmentation patterns in the zebrafish (Shoji *et al.*, 2002). We will here use reaction-diffusion models to show that the complex patterns of pigmentation in *R. imitator* can all be generated by varying the rate of diffusion of a single protein during development.

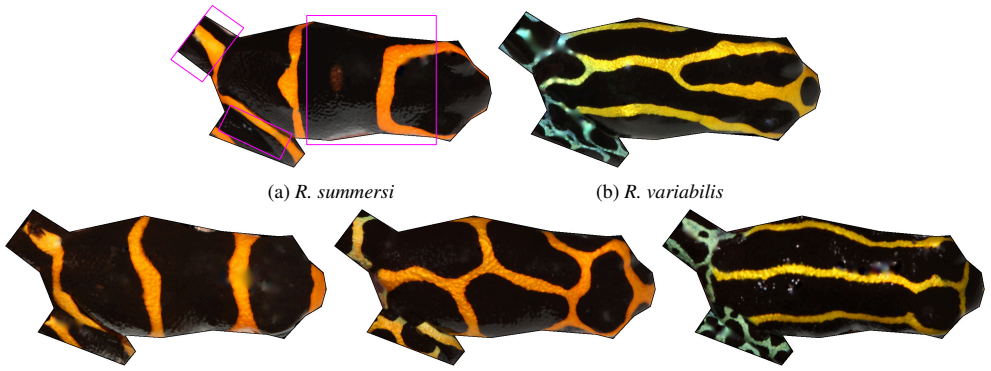
2. Quantifying phenotypes

Quantification of the phenotypes of interest is done in two steps: first, multiple features are extracted automatically from images using a suite of image descriptors and collected in a data $N \times p$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ where p is the number of features, and secondly, the p-dimensional feature space in which these quantities reside are reduced to a scalar z_i for each individual, representing the degree of mimicry with respect to this phentype. This dimensionality reduction will be described further below.

Phenotypes are extracted and reduced to a mimicry-related scalar for four separate case studies:

1. Dorsal saggital stripes vs. dorsal transversal stripes.
2. Dorsal pattern vs. dorsal coloration.
3. Dorsal pattern vs. leg pattern.
4. Dorsal coloration vs. leg coloration.

The first case is included as an example of a simple case, where we only use a single descriptor to quantify each of the two phenotypes. The distribution of pixel values for each row is used to calculate the entropy for each row and an average over these row entropies are used as a descriptor of saggital stripes; a minimum entropy is equivalent to all pixels having the same value, i.e., it is a saggital stripe. Conversely column sum entropies are averaged to obtain a descriptor directed at transversal stripes.



(a) *R. summersi* (b) *R. variabilis*

(c) *R. imitator* (Sauce) (d) *R. imitator* (Malpaso S. Side) (e) *R. imitator* (Micaela Bastidas)

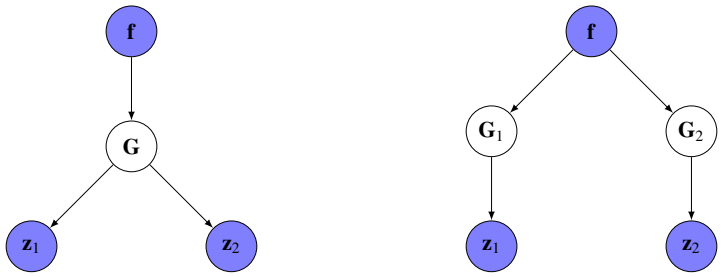
Figure 2: Illustration the two model species *R. summersi* and *R. variabilis* and the mimic frog *R. imitator*. The regions extracted as representatives for legs and dorsum are illustrated in (a).

Quantification of each of the remaining four phenotypes (dorsal pattern, leg pattern, dorsal coloration and leg coloration) are obtained using multivariate descriptors. Coloration is quantified by collecting all pixels in regions of interest in a hue-saturation-value (HSV) colorspace representation, performing K-means clustering with $K = 2$ and using the cluster center of the most populated cluster as a three-dimensional descriptor. Pattern is quantified as in Vestergaard *et al.* (2014) yielding a ten dimensional descriptor of pattern in the region of interest. The multivariate descriptors are reduced to one-dimensional mimicry-related phenotypic indices $\mathbf{z} = \mathbf{X}\mathbf{w}$, where $\mathbf{z} \in \mathbb{R}^N$ and $\mathbf{w} \in \mathbb{R}^p$ is determined using linear discriminant analysis (LDA) where the model species are used as training observations. The regularization parameter λ for LDA is chosen automatically as the minimal value of 50 values of $\{\lambda\}_i^{50}$ equidistantly spaced in the log-domain between $1e-5$ and $1e4$ for which the manifold was stable, i.e., where the average squared change over all individuals' phenotypic quantity when moving from λ_{i-1} to λ_i was below $1e-5$.

In each of the four cases, two separate data matrices are collected and $\mathbf{z}_1 \in \mathbb{R}^N$ and $\mathbf{z}_2 \in \mathbb{R}^N$ refer to the two quantified phenotypes for all N individuals.

3. Likelihood model

We will assume that we have samples from individuals from an admixture zone with a phenotype that is fixed at each end of the admixture zone. We will also assume that we for each individual, i, have obtained an estimate of the admixture fraction, f_i, and measurements of each of two phenotypes, z_{i1} and z_{i2}. The vectors of these observations for multiple individuals are denoted by $\mathbf{f}$, $\mathbf{z}_1$, and $\mathbf{z}_2$, and we wish to test if phenotype 1 and 2 are controlled by the same or different genes. To test this, we establish two models that quantify phenotypes as either conditionally independent given the genotype or completely independent as illustrated in Figure 3. The parameters of the models are $\boldsymbol{\theta} = [\mu_1^0, \mu_1^1, \mu_1^2, \sigma_1, \mu_2^0, \mu_2^1, \mu_2^2, \sigma_2]$, where μ_i^j is the mean phenotypic value for the jth genotype of the ith phenotype. σ_i is the variance for phenoype i. These models can then be used to construct likelihood ratios or Bayes factors.



(a) Model 1 (b) Model 2

Figure 3: The two likelihood scenarios modelled when observing the mixture proportions $\mathbf{f}$ and the phenotypes $\mathbf{z}_1$ and $\mathbf{z}_2$. Model 1 assumes that a single genetic component generate the two phenotypes, while model 2 assumes two conditionally independent genetic components when given the mixture proportions.

Assuming $z_1|G \sim \mathcal{N}(\mu_1^G, \sigma_1^2)$ and $z_2|G \sim \mathcal{N}(\mu_2^G, \sigma_2^2)$, the joint probability given the mixture proportions $\mathbf{f} = \{f_i\}_{i=1}^N$ can take two different forms:

$$\text{1. Same gene:} \quad p(\mathbf{z}_1, \mathbf{z}_2|\mathbf{f}; \boldsymbol{\theta}) = \prod_i^N \sum_{j\in\{0,1,2\}} p(z_{i1}|G=j)p(z_{i2}|G=j)p(G=j|f_i)$$

$$\text{2. Different genes:} \quad p(\mathbf{z}_1, \mathbf{z}_2|\mathbf{f}; \boldsymbol{\theta}) = \prod_i^N \left[\sum_{j\in\{0,1,2\}} p(z_{i1}|G=j)p(G=j|f_i)\right] \cdot \left[\sum_{j\in\{0,1,2\}} p(z_{i2}|G=j)p(G=j|f_i)\right]$$

with genotype probabilities

$$\begin{aligned} p(G=0|f) &= (1-f)^2 \\ p(G=1|f) &= 2f(1-f) \\ p(G=2|f) &= f^2 \,. \end{aligned} \tag{1}$$

Notice here that we have assumed that phenotypes are (conditionally) normal distributed with constant variance, and that there is a single di-allelic locus controlling the genotype. The genotype probabilities can be calculated from $\mathbf{f}$ as above, because of the previously mentioned assumption that the trait is fixed at each end of the admixture zone and is controlled by a single di-allelic locus. The models can be extended to allow for more than one locus. For K loci, the set of possible multi-locus genotypes is denoted $\mathcal{G}(K)$ with a multi-locus genotype denoted $\{\mathbf{g}_k\}_k^{|\mathcal{G}(K)|}$ where $\mathbf{g}_k = [g_1, \ldots, g_K]$ with $g_j \in \{0, 1, 2\}$. We now assume that

$$\begin{aligned} z_1|\mathbf{g}_k &\sim \mathcal{N}(\mathbf{h}_k^T \boldsymbol{\mu}_1, \sigma_1^2) \\ z_2|\mathbf{g}_k &\sim \mathcal{N}(\mathbf{h}_k^T \boldsymbol{\mu}_2, \sigma_2^2) \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_1 = [\mu_1^0, \mu_1^1, \mu_1^2]^T$, similarly for $\boldsymbol{\mu}_2$ and h_k is a three element vector holding proportions of the multi-locus genotype being either 0, 1 or 2, i.e., *AA*, *Aa* or *aa*. The likelihood models for either two traits controlled by the same K genes or two independent sets of K genes are then:

$$\text{1. Same genes:} \quad p(\mathbf{z}_1, \mathbf{z}_2|\mathbf{f}; \boldsymbol{\theta}) = \prod_i^N \sum_{\mathbf{g}_k \in \mathcal{G}(K)} p(z_{i1}|G = \mathbf{g}_k) p(z_{i2}|G = \mathbf{g}_k) p(G = \mathbf{g}_k|f_i)$$

$$\text{2. Different genes:} \quad p(\mathbf{z}_1, \mathbf{z}_2|\mathbf{f}; \boldsymbol{\theta}) = \prod_i^N \left[\sum_{\mathbf{g}_k \in \mathcal{G}(K)} p(z_{i1}|G = \mathbf{g}_k) p(G = \mathbf{g}_k|f_i) \right] \cdot \left[\sum_{\mathbf{g}_k \in \mathcal{G}(K)} p(z_{i2}|G = \mathbf{g}_k) p(G = \mathbf{g}_k|f_i) \right]$$

where $p(G = \mathbf{g}_k|f_i) = \prod_j^K p(g_j|f_i)$.

We also provide an alternative formulation that incorporates uncertainty in the estimates of f_i using a bootstrap approach, i.e. we assume that marker loci used for estimation of f_i have been bootstrapped to provide a bootstrap distribution $\mathbf{f}_i = \{f_i^b\}_{b=1}^B$:

$$p(G = \mathbf{g}_k|\mathbf{f}_i) = \frac{1}{B} \sum_{b=1}^{B} p(G = \mathbf{g}_k|f_i^b). \tag{3}$$

The two models have the same number of parameters. The likelihood ratio obtained by comparing the models can therefore be used directly as a measure of support for one model or another. No correction for degrees of freedom is necessary. For the purpose of hypothesis testing, we suggest bootstrapping individuals to provide a confidence region for the log likelihood ratio. If this region does not contain zero, one of the models can be rejected with statistical confidence.

4. Reaction-diffusion mechanism

Turing (1952) introduced reaction-diffusion (R-D) equations as a model for pattern formation: in the abscence of diffusion the concentrations of the two morphogens will stabilize, but under certain conditions diffusion driven instability can make spatially inhomogeneous patterns emerge (Murray, 2002). Gierer and Meinhardt (1972) have argued that local self-enhancement and long-range inhibition in the R-D system is the driving force of pattern formation and models have been derived on that basis for a wide range of biological systems, (see e.g., Bard, 1981; Meinhardt, 1993; Koch and Meinhardt, 1994; Meinhardt, 1999; Shoji *et al.*, 2003; Kondo and Miura, 2010; Allen *et al.*, 2013).

Here we consider the model by Turk (1991)

$$\frac{\partial A_{x,y}}{\partial t} = s(A_0 B_0 - A_{x,y} B_{x,y}) + D_a \alpha(\theta) \nabla^2 A_{x,y} \tag{4}$$

$$\frac{\partial B_{x,y}}{\partial t} = s(A_{x,y} B_{x,y} - B_{x,y} - \beta_{x,y}) + D_b \nabla^2 B_{x,y} \ , \tag{5}$$

with anisotropic diffusion $\alpha(\theta)$ as suggested by Shoji *et al.* (2003). The local diffusivity is governed by the Laplacian ∇^2, $A_{x,y}$ and $B_{x,y}$ are the concentrations of the two morphogens at time t at some position (x, y) in the domain with $A_{x,y} = A_0, B_{x,y} = B_0 \ \forall x, y$ at time $t = 0, \beta_{x,y} \sim \mathcal{N}(A_0 B_0 - B_0, \sigma_p^2)$ are small random perturbations, D_a and D_b are diffusivity coefficients for each morphogen, where $D_a > D_b$, and s is a scaling factor. The anisotropic diffusion weighting is defined as

$$\alpha(\theta) = \frac{1}{\sqrt{1 - \delta_a \cos 2\theta}} \tag{6}$$

with the anisotropy magnitude $\delta_a \in]-1, 1[$ and θ the angle between the two locations being considered. For $\delta_a \to -1$ prevalance is given to diffusion along the x-axis and along the y-axis for $\delta_a \to 1$.

To model the patterning of *R. imitator* in the studied transition zone we introduce a parameter $\kappa(x, y)$ distributed as a logistic function

$$\kappa(x, y) = \frac{1}{1 + e^{-\gamma(x - x_0)}} \ , \ \beta > 0, \kappa(x, y) \in [0, 1] \tag{7}$$

of the transversal position on the frog domain, where the position of the transition from legs to dorsum is x_0. β controls the steepness of the transition near $x = x_0$. Secondly, we model the magnitude of angular diffusion δ_a in Eq. (4) as a convex combination of two functions of the mixture proportion f:

$$\delta_a(x, y) = \kappa(x, y)\delta^{\text{dorsal}}(f) + (1 - \kappa(x, y))\delta^{\text{legs}}(f) \ , \ f \in [0, 1]$$

where $\delta^{\text{dorsal}}(f) = 0.9(2f - 1)$ and $\delta^{\text{legs}}(f) = 0.9(f - 1)$. This expression for the anisotropy magnitude is used in place of δ_a in Eq. (6). Hereby, we let the anisotropy magnitude vary spatially, with the spatial variation determined by the mixture proportion f; when $f = 0$ the anisotropy magnitude is $\delta_a(x, y) = -0.9$, i.e., the same over the entire domain, while at the other extreme $f = 1$ the anistropy magnitude $\delta_a(x, y) = 0.9\kappa(x, y)$, i.e., it varies according to the transversal position.

The simulations shown later have been carried out in a frog-like domain on a triangular mesh with 658 vertices. A sketch of the domain and the mapping functions defined above can be found in the supporting information S3 and $x_0 = 3$ is marked on the resulting patterns in Figure 8. The parameters used were $\gamma = 2$,$A_0 = B_0 = 4$, $\sigma_p = 0.001$, $s = 0.025$, $D_a = 0.175$, $D_b = 0.035$ and run for 1500 iterations.

5. Microsatellite data

We used published microsatellite data from two sources: Twomey *et al.* (2013) (92 samples), Twomey *et al.* (2014) (36 samples). In addition, we used 157 samples from an unpublished dataset (E. Twomey, J. S. Vestergaard and K. Summers, in preparation). The final microsatellite dataset consisted of 285 *R. imitator* individuals from 16 localities in Peru. For the unpublished microsatellite data, amplification methods follow Twomey *et al.* (2013).

We used JPEG compressed images of 6 *R.summersi*, 7 *R. variabilis* and 313 *R. imitator* individuals from 11 localities Both microsatellite data and image data were available for 179 of the *R. imitator* individuals.

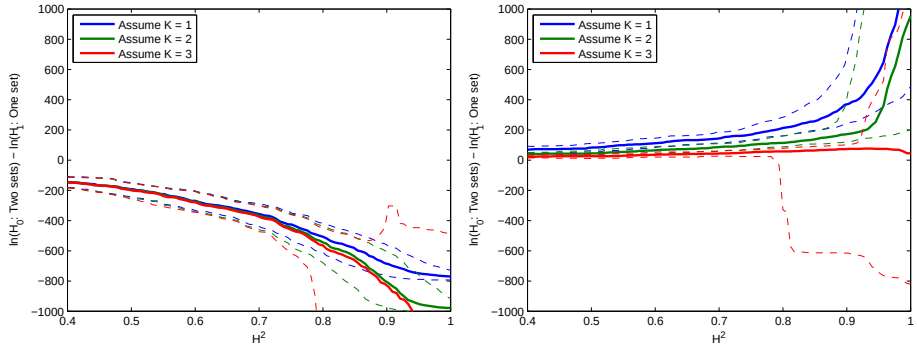
6. Results

6.1. Simulations

We evaluate the performance of the method using simulations allowing for varying heritability and uncertainty in the estimates of f. The heritability is modeled as in (Vestergaard *et al.*, 2014). To simulate data for two phenotypes determined by the same set of K genes, n mixture proportions $\mathbf{f} = \{f_i\}_1^n$ are drawn from a uniform distribution on the interval [0,1]. The genotype for each of the K loci are then drawn from a multinomial distribution with probabilities as in Eq. (1). To simulate data for two phenotypes determined by two separate sets of genes, the genotype for each of the K loci are drawn independently for the two phenotypes. Phenotypes are then assigned by simulating from a normal distribution as in Equation (2). In simulations with noise in the estimate of f we simulate $B = 100$ samples from a normal distribution with standard deviation σ_f around f_i, such that admixture proportions used for inference are distributed as $\hat{f}_i^b \sim \mathcal{N}\left(f_i, \sigma_f^2\right)$.

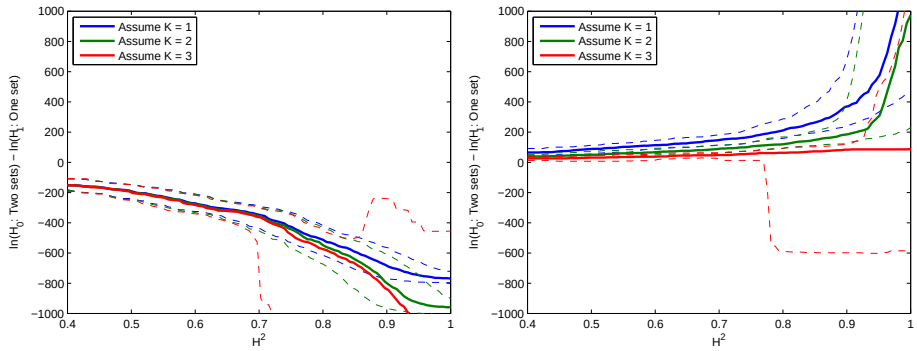
The simulation studies (Figure 4) show that the correct model is always identified when simulating data under the same model used for inference. This is true both with and without noise in the estimate of $\mathbf{f}$. However, when the true model includes two separate sets of genes, fitting a model with K larger than the true number of genes can lead to erroneous estimates (i.e. a likelihood ratio favoring a model with both phenotypes controlled by the same loci). We also note that the opposite is not true. Assuming fewer loci than the true number of loci does not lead to false inferences. In fact, in all simulations, assuming $K = 1$ leads to a likelihood ratio that always supports the true model. Thus, we recommend fitting a model with $K = 1$ genes when doing inference under these models. This choice is robust even when $K > 1$. Additional simulation studies can be found in the supporting information S1.

Figure 5 shows receiver-operator characteristic (ROC) curves for three different sample sizes $N = \{100, 250, 1000\}$, two heritability coefficients $H^2 = \{0.5, 0.9\}$ for simulated data under $K = \{1, 3\}$ genes and three different noise levels on the mixture proportions. 1000 simulations were used for each ROC curve, with 500 replicates assuming each of two models being the true model. A true positive is characterized as correctly inferring model 1, while a false positive is inferring model 1 when the true model is model 2 (as determined by the likelihood ratio). The ROC curves show that for sample sizes $N > 100$ near-perfect inference is possible, while for small sample sizes of $N = 100$ and heritability $H^2 = 0.5$ a slight reduction in performance is observed.



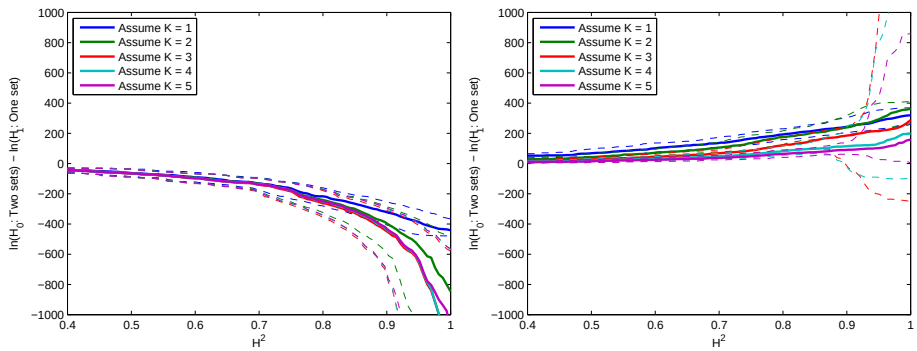
(a) One set, $K = 1, f \sim \mathcal{U}(0, 1)$

(b) Two sets, $K = 1, f \sim \mathcal{U}(0, 1)$



(c) One set, $K = 1, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.05^2)$

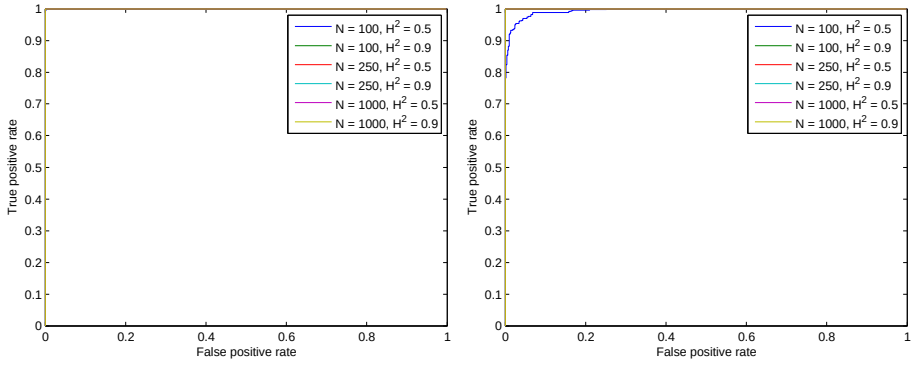
(d) Two sets, $K = 1, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.05^2)$



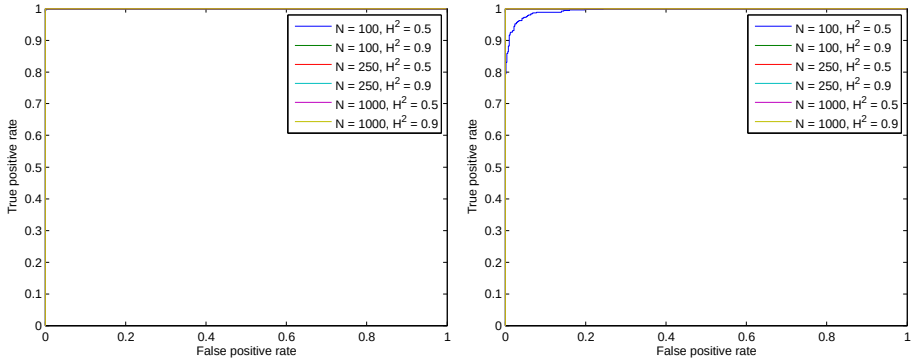
(e) One set, $K = 3, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.1^2)$

(f) Two sets, $K = 3, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.1^2)$

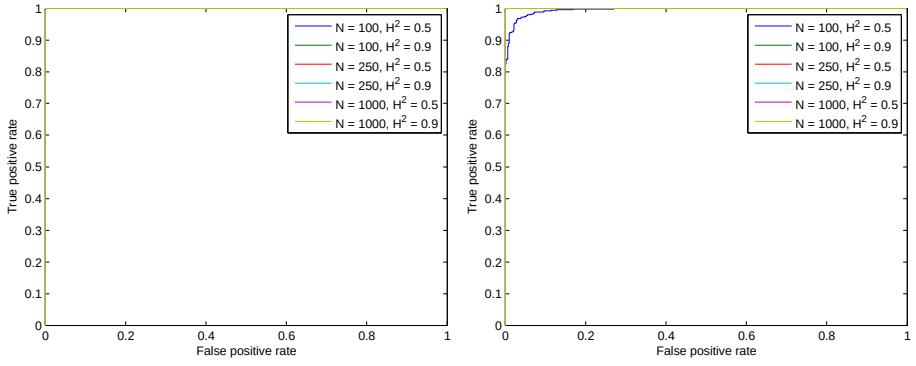
Figure 4: Log-likelihood ratios for simulation studies. The solid curves are medians based on 500 simulations smoothed with a Gaussian kernel with standard deviation 0.05 and the dashed lines are 5 and 95 percentiles. The true parameters used to simulate the data are shown in the captions.



(a) $K = 1, f \sim \mathcal{U}(0, 1)$ (b) $K = 3, f \sim \mathcal{U}(0, 1)$



(c) $K = 1, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.05^2)$ (d) $K = 3, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.05^2)$

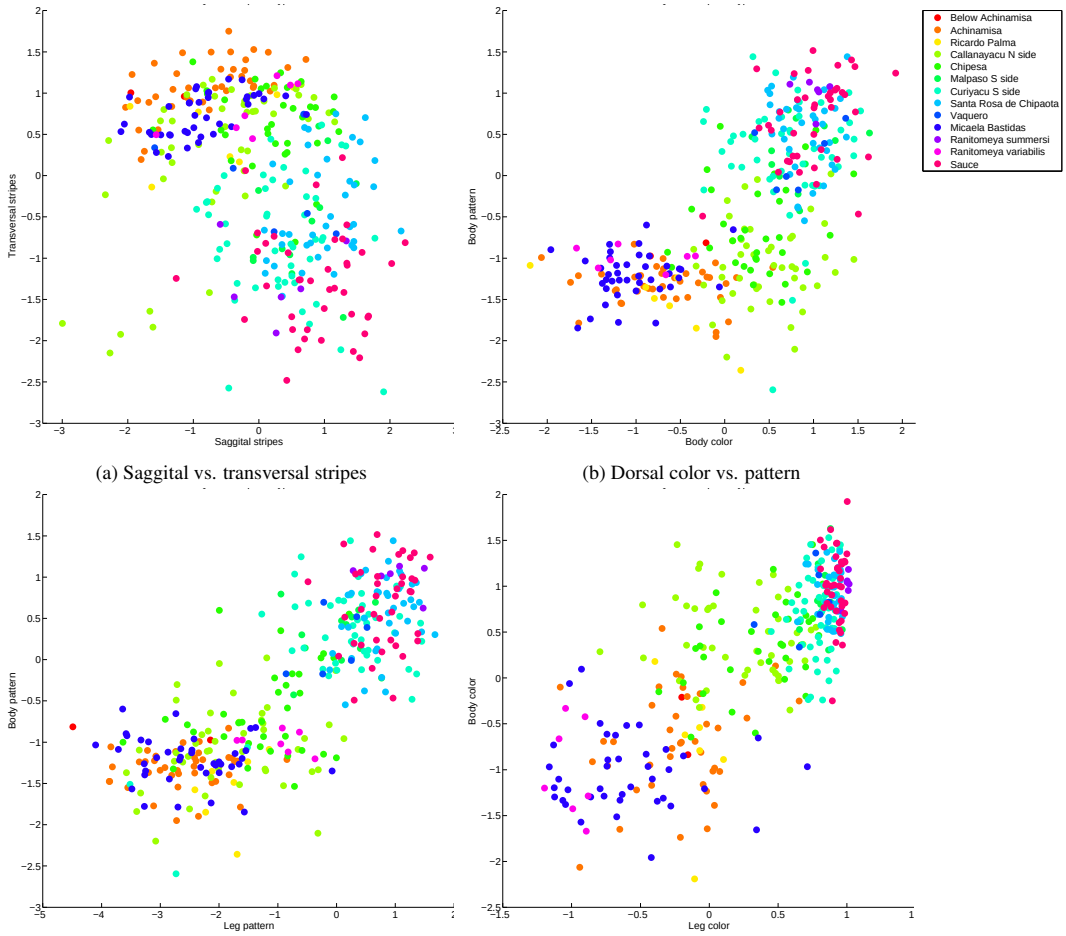


(e) $K = 1, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.1^2)$ (f) $K = 3, f \sim \mathcal{U}(0, 1) + \mathcal{N}(0, 0.1^2)$

Figure 5: Receiver-operator characteristic (ROC) curves for $N = \{100, 250, 1000\}, K = \{1, 3\}, \sigma_f = \{0, 0.05, 0.1\}$ and two heritabilities $H^2 = \{0.5, 0.9\}$. $B = 100$ are used in all simulations.

6.2. *Case studies*

The quantified mimicry-related phenotypes are shown as scatter plots in Figure 6. Individuals are colored according to location and to some extent locations cluster together in this two-dimensional space. Note for instance the simple case in Figure ((a)) that *R. imitator* individuals from Sauce are found in the proximity of *R. summersi* individuals, *R. imitator* individuals from Achinamisa are found in the proximity of *R. variabilis* individuals, and in the middle are *R. imitator* individuals from Malpaso South Side and Curiyacu South Side. This means that, despite the simplicity of the extracted phenotype, there is a lot of meaningful signal captured. Similar inspections can be made from the other scatter plots, but the complexity of the extracted phenotypes and subsequent dimensionality reduction makes the interpretation more difficult.



(a) Saggital vs. transversal stripes

(b) Dorsal color vs. pattern

(c) Leg vs. dorsal pattern

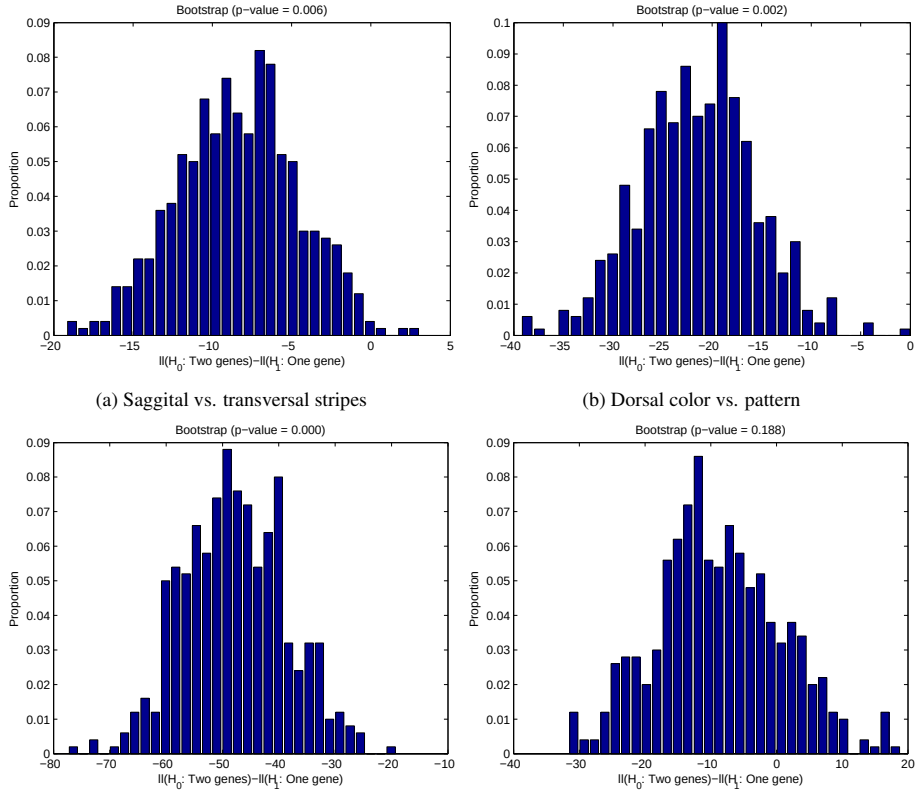
(d) Leg vs. dorsal color

Figure 6: Scatter plots of the two quantified phenotypes for the four cases.

The likelihood models presented above were fitted to each of these four cases. In each case, the mixture proportions were bootstrapped $B = 100$ times and likelihood ratios were bootstrapped 500 times. The full bootstrap distribution of likelihood ratios associated with the hypothesis of H_0 : 2 genes against H_A : 1 gene is shown in Figure 7 for the four cases. Supporting information S2 contains similar plots when assuming $K = 2$ and $K = 3$.

The p-values associated with each case study, assuming one of $K = \{1, 2, 3\}$ genes are shown in Table 1. Note that we estimate that saggital stripes and transversal stripes are controlled by the same underlying gene, the same goes for dorsal color and pattern, and for dorsal pattern and leg pattern. However, there is no evidence in favor of

the hypothesis that leg color and dorsal color are controlled by the same underlying gene. These conclusions are true for $K = \{1, 2, 3\}$ genes, and do generally not seem to depend much on assumptions regarding K.



(a) Saggital vs. transversal stripes

(b) Dorsal color vs. pattern

(c) Leg vs. dorsal pattern

(d) Leg vs. dorsal color

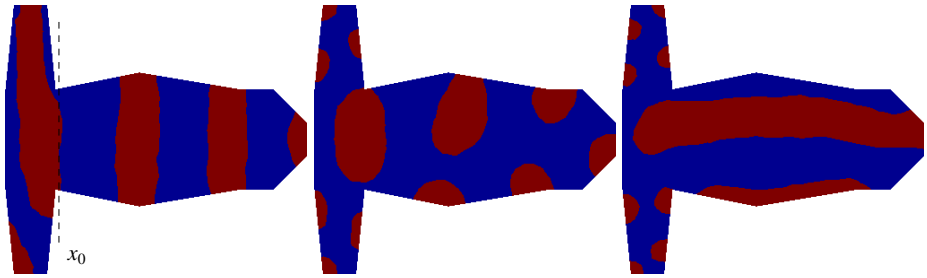
Figure 7: Bootstrap distribution of log-likelihood ratios for the null hypothesis of selecting a model with two separate sets of underlying genes against the alternative of selecting a model with a single set of underlying genes. 500 bootstrap samples were used.

Phenotype 1	Phenotype 2	K = 1	K = 2	K = 3
Saggital stripes	Transversal stripes	**0.006**	**0.010**	**0.032**
Body color	Body pattern	**0.002**	**0.002**	**0.000**
Leg pattern	Body pattern	**0.000**	**0.000**	**0.000**
Leg color	Body color	0.188	0.280	0.164

Table 1: P-values for the null hypothesis of two independent sets of K genes versus the alternative of a single set of K genes underlying the two phenotypes. P-values below 0.05 are marked in bold.

6.3. Reaction-diffusion simulations

To illustrate that a single gene can in fact control two phenotypes as complex as dorsal pattern and leg pattern, the reaction-diffusion model described earlier was run for $f = 0$ corresponding to individuals mimicking *R. summersi*, $f = 0.4$ corresponding to the individual in Figure 2(d) and $f = 1$ corresponding to an individual mimicking *R. variabilis*. The results of these simulations can be seen in Figure 8 as a binary image with values of 1 (red) being areas where $A > A_0$ and 0 (blue) being areas where $A \leq A_0$ after convergence. While these simulations do not perfectly replicate the complex patterns of *R. imitator*, they do serve as an illustration of what a simple one-parameter mathematical model can achieve.



(a) $f = 0$ (b) $f = 0.4$ (c) $f = 1$

Figure 8: Patterns generated by the reaction-diffusion model. The parameter f controls the synthesized admixture proportion, where $f = 0$ or $f = 1$ corresponds to the two model species respectively. The admixture proportion of $f = 0.4$ is equal to the estimated admixture proportion for the *R. imitator* from Malpaso S. Side in Figure 2. x_0 refers to the parameter in Eq. (7).

7. Discussion

We have developed a statistical approach to determine whether two phenotypes are controlled by the same set of genes, or by multiple sets of genes, using individuals sampled from an admixture (or introgression) zone. While this approach may be superfluous for organisms that can easily be bred in captivity, it should be useful for research on species that do not easily lend themselves to captive breeding. We were able to show that a single causal gene is a relatively conservative assumption that should not lead to false inferences, and therefore recommend the use of this model for inferences. We applied the method to data from *R. imitator* and were able to show that all major phenotypes could potentially be controlled by the same set of genes, with the exception of leg color which may be controlled by a different set.

While it does indeed seem remarkable that a single set of genes could control this diversity of phenotypes, the simplicity of the presented reaction-diffusion model illustrates that fairly simple mechanisms can generate very diverse phenotypes, such as the patterns characteristic of of *R. imitator* morphs. Based on these results, the manipulation of a single parameter in the reaction-diffusion model can reasonably account for three of the four known mimetic morphs. For example, the banded morph, at least in the "pure" mimetic populations, has two major crossbands on the dorsum, one small crossband on the nose, and one running transverse across the dorsal surfaces of the legs, a pattern that is nearly perfectly recovered when the synthesized admixture proportion is set equal to zero (Figure 8(a)). Increasing the admixture proportion, the spotted morph is reasonably well approximated (Figure 8(b)), with irregular spots on the dorsum and reticulated legs. Finally, the striped morph is loosely approximated (Figure 8(c)) by increasing the synthesized admixture proportion to one, with lengthwise dorsal stripes and reticulated legs.

Further, the results are consistent with field observations. Based on our field work, there appear to be two major axes of color variation: dorsal/head coloration (including ventral side of chin), and limb/ventral coloration. Across all populations, dorsal/head coloration varies from green to yellow to orange, and leg/ventral coloration varies from greenish blue to navy blue to orange in the banded population. Most permutations of the above dorsal and leg colors seem possible. For example, we have observed striped-pattern and reticulated-pattern frogs with green, yellow, or orange dorsal coloration, and a wide variety of leg coloration as well. The major exception is the banded morph, in which the presence of latitudinal dorsal bands is invariably accompanied with orange dorsal and leg coloration, and a lack of reticulation on the legs.

These results may provide insight into other systems that show similar combinations of color pattern across distinct body regions. While aposematism and mimicry have attracted substantial interest and stimulated much research in evolutionary biology, there have as yet been few overall syntheses that investigate correlations (or lack thereof) in the mechanisms that underlie similarities and differences in color pattern between populations, species and higher taxa. Hence the general question, "Do similar or diverse mechanisms underlie the development of aposematic and mimetic coloration in different organisms across the tree of life?", remains open. Our method for estimating genetic control of color pattern will allow researchers to estimate this important feature of genetic control in the many organisms where the production of extensive lab crosses is not feasible. Further, our reaction-diffusion modeling approach may provide a heuristic tool with which researchers can get an initial sense of whether distinct suites of color pattern across populations, species or higher taxa are likely to be controlled by similar genetic factors and developmental rules.

The statistical methodology developed in this paper allows researchers to determine if several phenotypic traits are controlled by the same set, or by multiple different sets, of genes. This is an important approach for understanding the genetic basis of adaptive traits. However, it is also important in studies aimed at directly mapping the genetic basis of the trait, for example using association mapping or divergence mapping. If multiple traits are controlled by the same gene this significantly simplifies the mapping procedures that can combine measurements from multiple phenotypes to provide a more accurate description of a composite phenotype (see e.g., Vestergaard *et al.*, 2014). It also reduces the multiple testing burden. The methodology described here should, therefore, find general use in studies aimed at understanding the genetic basis of adaptive traits.

Publicly available implementations of the presented methods can be found at `https://github.com/schackv`.

References

Allen, W. L., R. Baddeley, N. E. Scott-Samuel, and I. C. Cuthill, 2013 The evolution and function of pattern diversity in snakes. Behavioral Ecology *24*(5)**:** 1237–1250.

Bard, J. B., 1981 A model for generating aspects of zebra and other mammalian coat patterns. Journal of Theoretical Biology *93*(2)**:** 363–85.

Baxter, S. W., N. J. Nadeau, L. S. Maroja, P. Wilkinson, B. A. Counterman, A. Dawson, M. Beltran, S. Perez-Espona, N. Chamberlain, L. Ferguson, and Others, 2010 Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the Heliconius melpomene clade. PLoS Genetics *6*(2)**:** e1000794.

Chouteau, M., K. Summers, V. Morales, and B. Angers, 2011 Advergence in Müllerian mimicry: the case of the poison dart frogs of Northern Peru revisited. Biology letters *7*(5)**:** 796–800.

Counterman, B. A., F. Araujo-Perez, H. M. Hines, S. W. Baxter, C. M. Morrison, D. P. Lindstrom, R. Papa, L. Ferguson, M. Joron, C. P. Smith, and Others, 2010 Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in Heliconius erato. PLoS Genetics *6*(2)**:** e1000796.

Gierer, a. and H. Meinhardt, 1972 A theory of biological pattern formation. Kybernetik *12*(1)**:** 30–9.

Joron, M., L. Frezal, R. T. Jones, N. L. Chamberlain, S. F. Lee, C. R. Haag, A. Whibley, M. Becuwe, S. W. Baxter, and L. Ferguson, 2011 Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature *477*(7363)**:** 203–206.

Joron, M. and J. L. Mallet, 1998 Diversity in mimicry: paradox or paradigm? Trends in Ecology & Evolution *13*(11)**:** 461–466.

Koch, A. and H. Meinhardt, 1994 Biological pattern formation: from basic mechanisms to complex structures. Reviews of Modern Physics *66*(4)**:** 1481.

Kondo, S. and T. Miura, 2010 Reaction-diffusion model as a framework for understanding biological pattern formation. Science (New York, N.Y.) *329*(5999)**:** 1616–20.

Kunte, K., W. Zhang, A. Tenger-Trolander, D. Palmer, A. Martin, R. Reed, S. Mullen, and M. Kronforst, 2014 doublesex is a mimicry supergene. Nature *507*(7491)**:** 229–232.

Martin, A. and V. Orgogozo, 2013 The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution *67*(5)**:** 1235–1250.

Martin, A., R. Papa, N. J. Nadeau, R. I. Hill, B. A. Counterman, G. Halder, C. D. Jiggins, M. R. Kronforst, A. D. Long, W. O. McMillan, and Others, 2012 Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. Proceedings of the National Academy of Sciences *109*(31)**:** 12632–12637.

Meinhardt, H., 1982 *Models of biological pattern formation*. Academic Press London.

Meinhardt, H., 1993 A model for pattern formation of hypostome, tentacles, and foot in hydra: how to form structures close to each other, how to form them at a distance. Developmental biology *157*(2)**:** 321–333.

Meinhardt, H., 1999 Orientation of chemotactic cells and growth cones: models and mechanisms. Journal of Cell Science *112*(17)**:** 2867–2874.

Murray, J., 2002 *Mathematical biology* (3rd editio ed.). Springer.

Nachman, M. W., H. E. Hoekstra, and S. L. D'Agostino, 2003 The genetic basis of adaptive melanism in pocket mice. Proceedings of the National Academy of Sciences *100*(9)**:** 5268–5273.

Papa, R., A. Martin, and R. D. Reed, 2008 Genomic hotspots of adaptation in butterfly wing pattern evolution. Current opinion in genetics & development *18*(6)**:** 559–564.

Sheppard, P. M., J. R. G. Turner, K. S. Brown, W. W. Benson, and M. C. Singer, 1985 Genetics and the evolution of Muellerian mimicry in Heliconius butterflies. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences**:** 433–610.

Shoji, H., Y. Iwasa, A. Mochizuki, and S. Kondo, 2002 Directionality of stripes formed by anisotropic reaction-diffusion models. Journal of Theoretical Biology *214*(4)**:** 549–61.

Shoji, H., A. Mochizuki, Y. Iwasa, M. Hirata, T. Watanabe, S. Hioki, and S. Kondo, 2003 Origin of directionality in the fish stripe pattern. Developmental dynamics : an official publication of the American Association of Anatomists *226*(4)**:** 627–33.

Symula, R., R. Schulte, and K. Summers, 2001 Molecular phylogenetic evidence for a mimetic radiation in Peruvian poison frogs supports a Müllerian mimicry hypothesis. Proceedings of the Royal Society of London. Series B: Biological Sciences *268*(1484)**:** 2415–2421.

Symula, R., R. Schulte, and K. Summers, 2003 Molecular systematics and phylogeography of Amazonian poison frogs of the genus *Dendrobates*. Molecular Phylogenetics and Evolution *26*(3)**:** 452–475.

Turing, A., 1952 The Chemical Basis of Morphogenesis. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences *237*(641)**:** 37–72.

Turk, G., 1991 Generating textures on arbitrary surfaces using reaction-diffusion. ACM SIGGRAPH Computer Graphics *25*(4)**:** 289–298.

Twomey, E., J. S. Vestergaard, and K. Summers, 2014 Reproductive isolation related to mimetic divergence in the poison frog Ranitomeya imitator. accepted for Nature Communications.

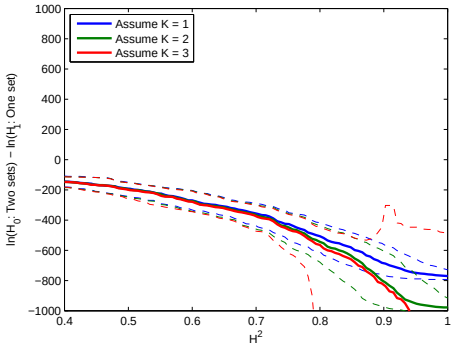
Twomey, E., J. Yeager, and J. Brown, 2013 Phenotypic and Genetic Divergence among Poison Frog Populations in a Mimetic Radiation. PloS One *8*(2).

Vestergaard, J. S., E. Twomey, R. Larsen, K. Summers, and R. Nielsen, 2014 Number of genes controlling a quantitative trait in a hybrid zone of the aposematic frog Ranitomeya imitator. Proceedings of the Royal Society. Series B. Biological Sciences [in review].

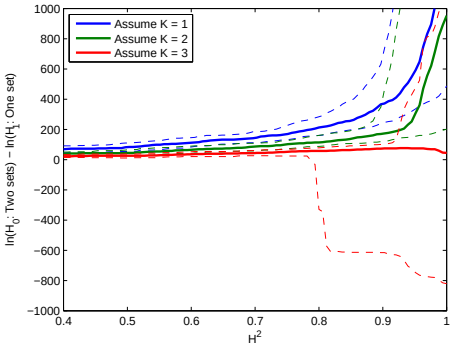
Yeager, J., J. L. Brown, V. Morales, M. Cummings, and K. Summers, 2012 Testing for selection on color and pattern in a mimetic radiation. Current Zoology **58:** 668–676.

File S1 Additional simulation studies

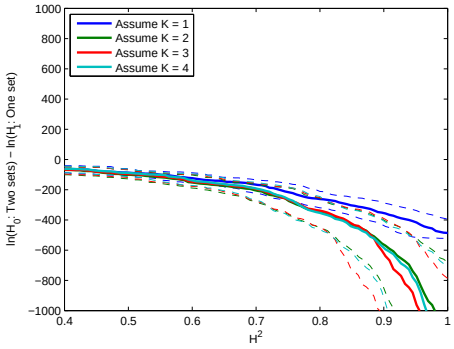
Log-likelihood ratios for additional simulation studies. The solid curves are medians based on 500 simulations smoothed with a Gaussian kernel with standard deviation 0.05 and the dashed lines are 5 and 95 percentiles. The true parameters used to simulate the data are shown in the captions.



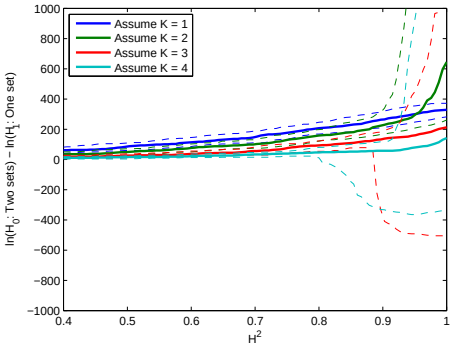
(a) One set, $K = 1, f \sim \mathcal{U}(0, 1)$



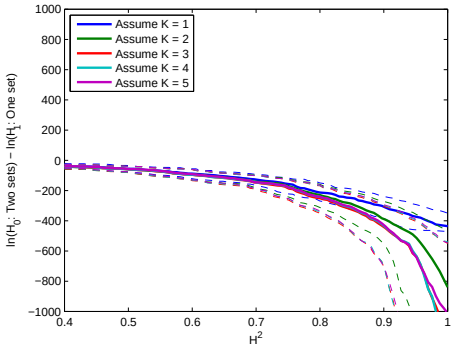
(b) Two sets, $K = 1, f \sim \mathcal{U}(0, 1)$



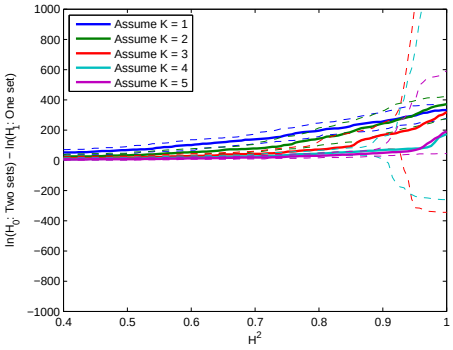
(c) One set, $K = 2, f \sim \mathcal{U}(0, 1)$



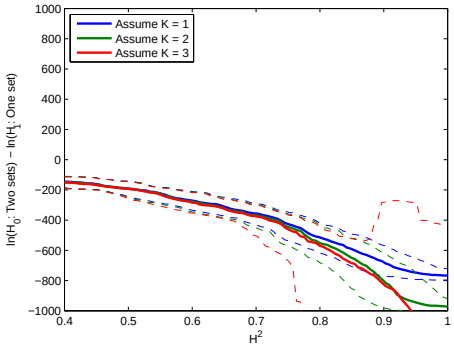
(d) Two sets, $K = 2, f \sim \mathcal{U}(0, 1)$



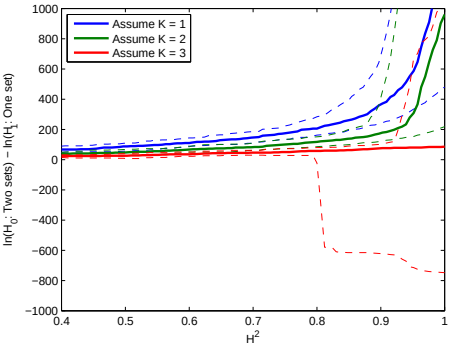
(e) One set, $K = 3, f \sim \mathcal{U}(0, 1)$



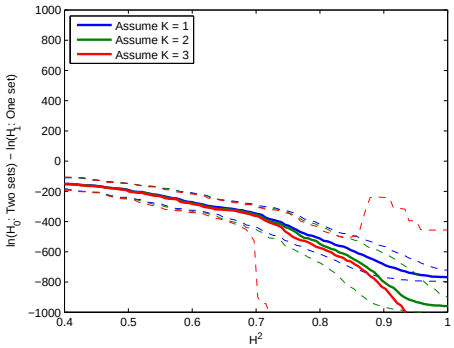
(f) Two sets, $K = 3, f \sim \mathcal{U}(0, 1)$



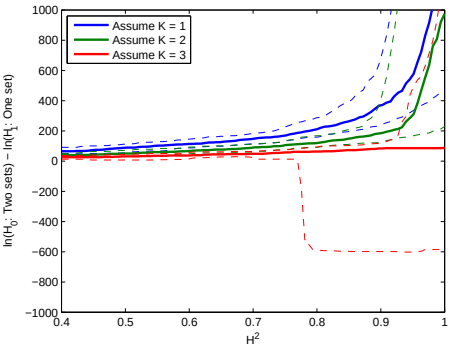
(g) One set, $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.01^2)$



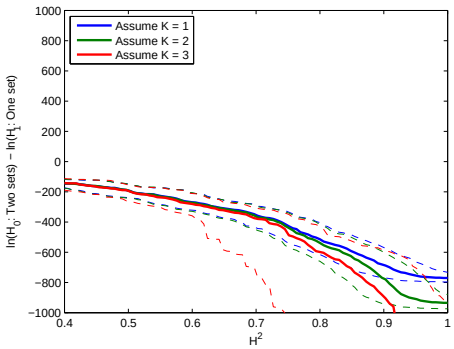
(h) Two sets, $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.01^2)$



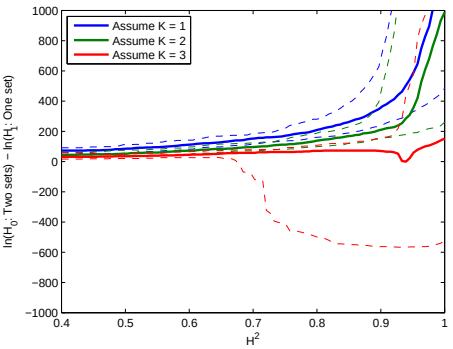
(i) One set, $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.05^2)$



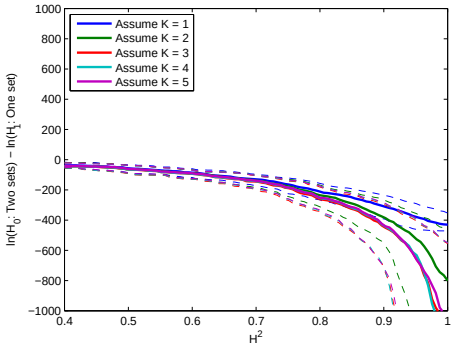
(j) Two sets, $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.05^2)$



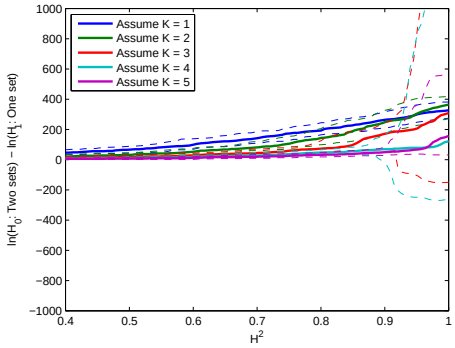
(k) One set, $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.1^2)$



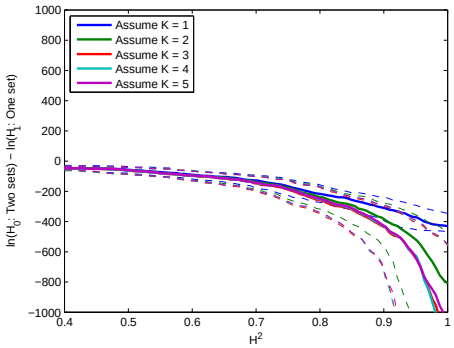
(l) Two sets, $K = 1, f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.1^2)$



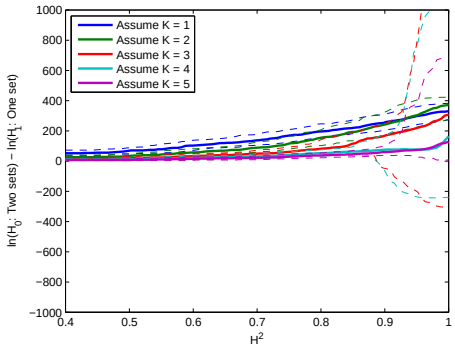
(m) One set, $K = 3$, $f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.01^2)$



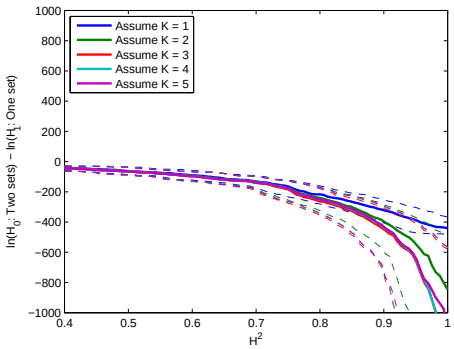
(n) Two sets, $K = 3$, $f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.01^2)$



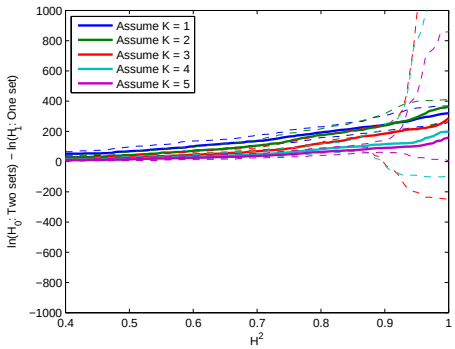
(o) One set, $K = 3$, $f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.05^2)$



(p) Two sets, $K = 3$, $f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.05^2)$



(q) One set, $K = 3$, $f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.1^2)$



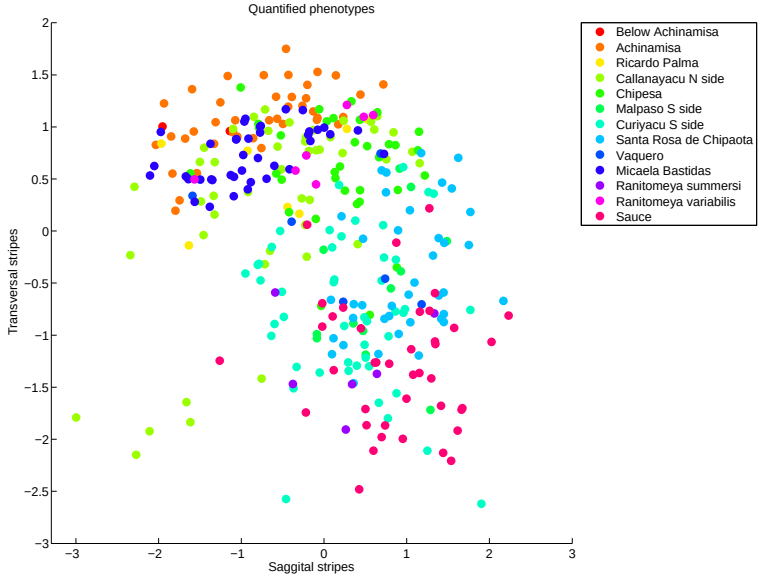
(r) Two sets, $K = 3$, $f \sim \mathcal{U}(0,1) + \mathcal{N}(0, 0.1^2)$

File S2 Case studies — auxilliary plots

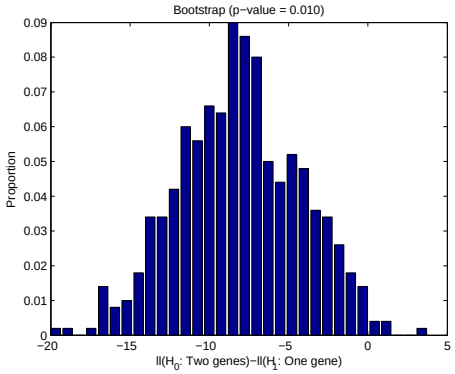
This supporting information provides auxilliary visualizations for each of the four cases treated in the main text. Specifically, three extra plots and a table is provided for each case: first a two-dimensional scatter plot of the two quantified phenotypes, where each observation is colored according to sample location. Secondly, bootstrap distributions of the hypothesis of the same versus two sets of genes for $K = 2$ and $K = 3$ genes in each set. The bootstrap p-values for rejecting the null hypothesis of two separate sets of genes are shown as plot titles. The bootstrap distribution for this hypothesis where $K = 1$ can be found in the main text. The included table provides the maximum likelihood point estimates of the parameters for $K = \{1, 2, 3\}$ for each of the two hypotheses.

Each case is presented on a single of the next pages in the following order: Saggital vs. transversal stripes, dorsal color vs. pattern, leg vs. dorsal pattern and leg vs. dorsal color.

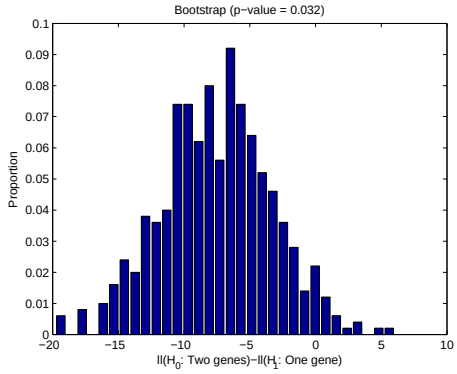
Saggital vs. transversal stripes



(a) Quantified phenotypes. Color codes correspond to locations shown in the legend.



(b) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 2$.

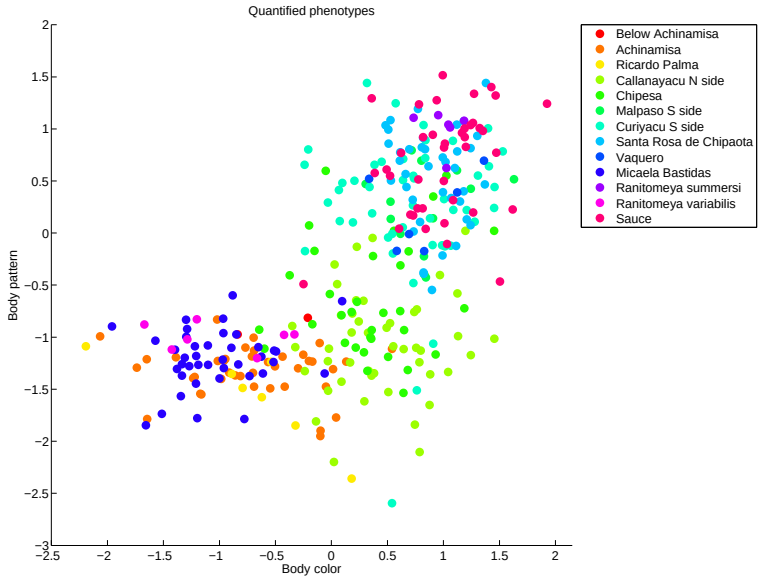


(c) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 3$.

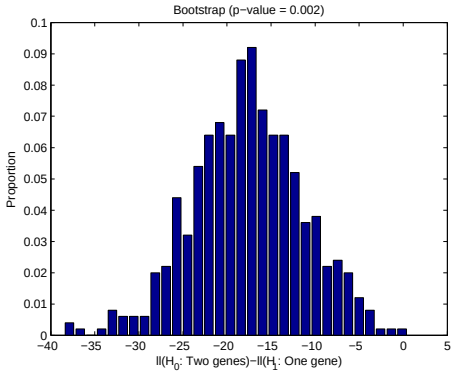
	One set	Two sets
$K = 1$,	$[-1.14, 0.35, 0.66, 0.61, 0.59, 0.55, -1.19, 0.49]$,	$[-1.19, 0.09, 0.76, 0.57, 0.66, 0.48, -1.18, 0.47]$
$K = 2$,	$[-1.26, 0.45, 0.70, 0.63, 0.64, 0.80, -1.48, 0.36]$,	$[-1.26, 0.04, 0.82, 0.61, 0.67, 0.77, -1.47, 0.36]$
$K = 3$:	$[-1.26, 0.40, 0.73, 0.68, 0.65, 0.94, -1.61, 0.31]$,	$[-1.22, -0.11, 0.85, 0.66, 0.69, 0.91, -1.61, 0.31]$

Table 1: Point estimates $[\mu_1^0, \mu_1^1, \mu_1^2, \sigma_1, \mu_2^0, \mu_2^1, \mu_2^2, \sigma_2]$

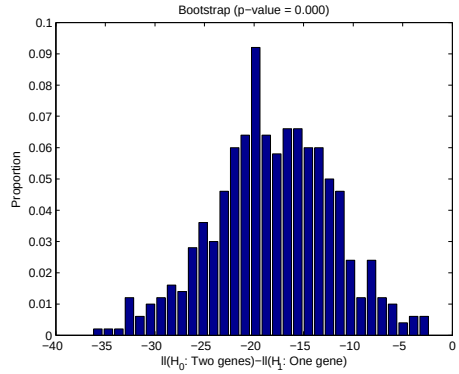
Dorsal color vs. pattern



(a) Quantified phenotypes. Color codes correspond to locations shown in the legend.



(b) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 2$.

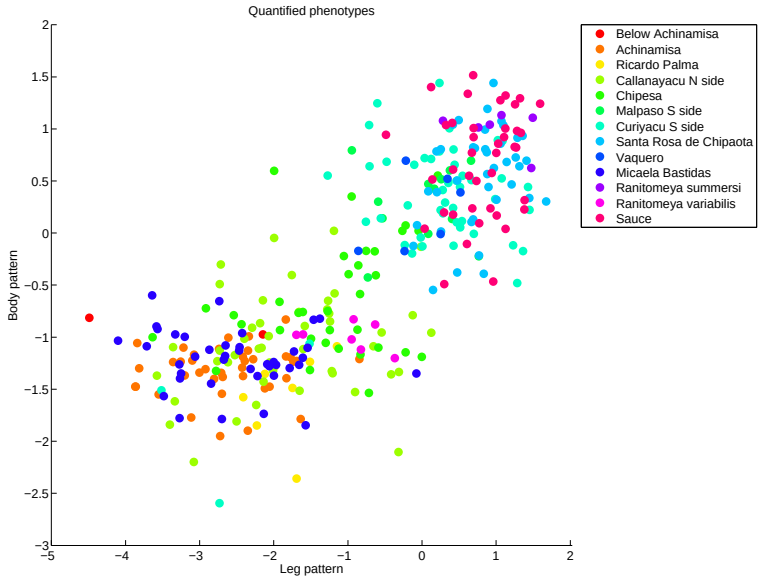


(c) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 3$.

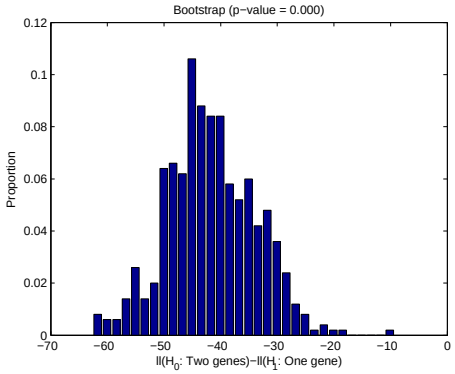
	One set	Two sets
$K = 1$,	$[-0.91, 0.50, 0.87, 0.45, -1.23, -0.89, 0.43, 0.40]$,	$[-0.96, 0.36, 0.95, 0.41, -1.17, -0.88, 0.44, 0.41]$
$K = 2$,	$[-0.99, 0.64, 0.95, 0.42, -1.21, -1.05, 0.64, 0.37]$,	$[-1.02, 0.58, 1.01, 0.39, -1.16, -1.10, 0.65, 0.36]$
$K = 3$:	$[-1.04, 0.71, 0.98, 0.41, -1.20, -1.17, 0.69, 0.37]$,	$[-1.06, 0.69, 1.03, 0.38, -1.14, -1.27, 0.69, 0.36]$

Table 2: Point estimates $[\mu_1^0, \mu_1^1, \mu_1^2, \sigma_1, \mu_2^0, \mu_2^1, \mu_2^2, \sigma_2]$

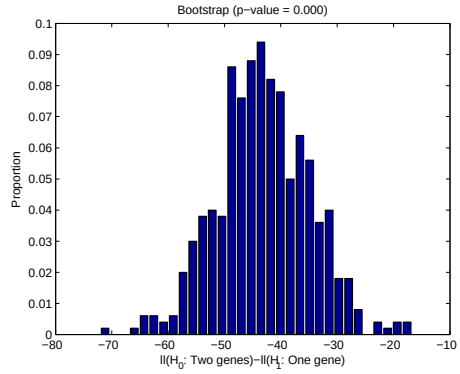
Leg vs. dorsal pattern



(a) Quantified phenotypes. Color codes correspond to locations shown in the legend.



(b) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 2$.

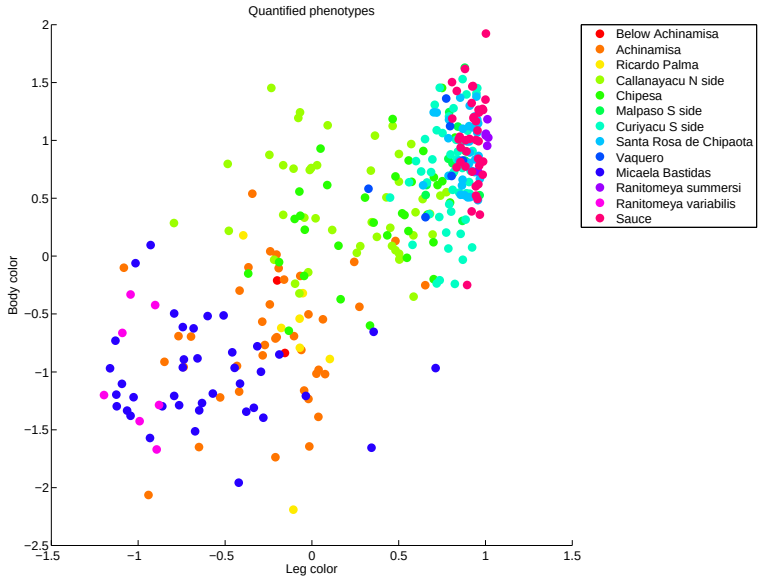


(c) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 3$.

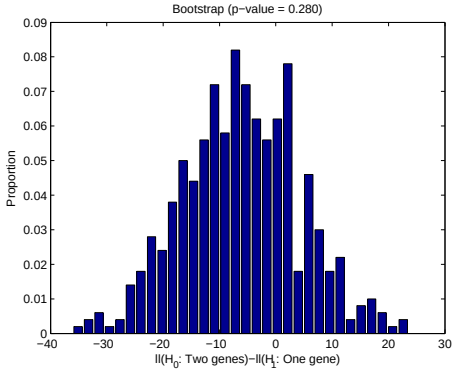
	One set	Two sets
$K = 1$,	$[-2.65, -1.27, 0.44, 0.68, -1.16, -0.95, 0.41, 0.42]$,	$[-2.66, -1.26, 0.48, 0.65, -1.17, -0.88, 0.44, 0.41]$
$K = 2$,	$[-2.82, -0.95, 0.65, 0.65, -1.16, -1.02, 0.63, 0.38]$,	$[-2.82, -1.07, 0.70, 0.62, -1.16, -1.11, 0.65, 0.36]$
$K = 3$:	$[-2.80, -1.32, 0.72, 0.68, -1.13, -1.27, 0.66, 0.37]$,	$[-2.84, -1.21, 0.76, 0.66, -1.14, -1.28, 0.69, 0.36]$

Table 3: Point estimates $[\mu_1^0, \mu_1^1, \mu_1^2, \sigma_1, \mu_2^0, \mu_2^1, \mu_2^2, \sigma_2]$

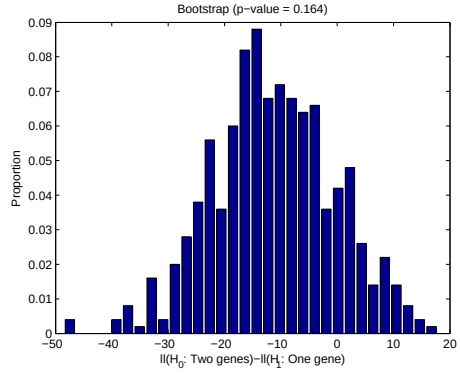
Leg vs. dorsal color



(a) Quantified phenotypes. Color codes correspond to locations shown in the legend.



(b) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 2$.



(c) Log-likelihood ratios $p_{\frac{H_0}{H_1}}$ given $K = 3$.

	One set	Two sets
$K = 1$,	$[-0.69, -0.00, 0.78, 0.21, -0.97, 0.11, 0.77, 0.53]$,	$[-0.75, -0.06, 0.77, 0.18, -0.95, 0.37, 0.96, 0.41]$
$K = 2$,	$[-0.79, 0.59, 0.86, 0.16, -0.88, 0.31, 1.01, 0.54]$,	$[-0.81, 0.60, 0.84, 0.16, -1.03, 0.56, 0.99, 0.39]$
$K = 3$:	$[-0.78, 0.63, 0.86, 0.18, -0.93, 0.46, 1.00, 0.52]$,	$[-0.79, 0.62, 0.86, 0.17, -1.06, 0.68, 1.03, 0.38]$

Table 4: Point estimates $[\mu_1^0, \mu_1^1, \mu_1^2, \sigma_1, \mu_2^0, \mu_2^1, \mu_2^2, \sigma_2]$

File S3 Reaction-diffusion model — illustrations

The presented reaction-diffusion model relies on a pre-defined spatial domain and two functions that vary depending on the position in this domain. Here we show illustrations that may be useful to help under understand the model presented in the main text. This includes the spatial domain in which the simulation is carried out, the function distributing the preparatory parameter over this domain and the two linear functions depending on the mixture proportion f.

This is a supplement for the main article and is thus not self-contained, i.e., the reaction-diffusion model is not restated here but can be found in the main text.

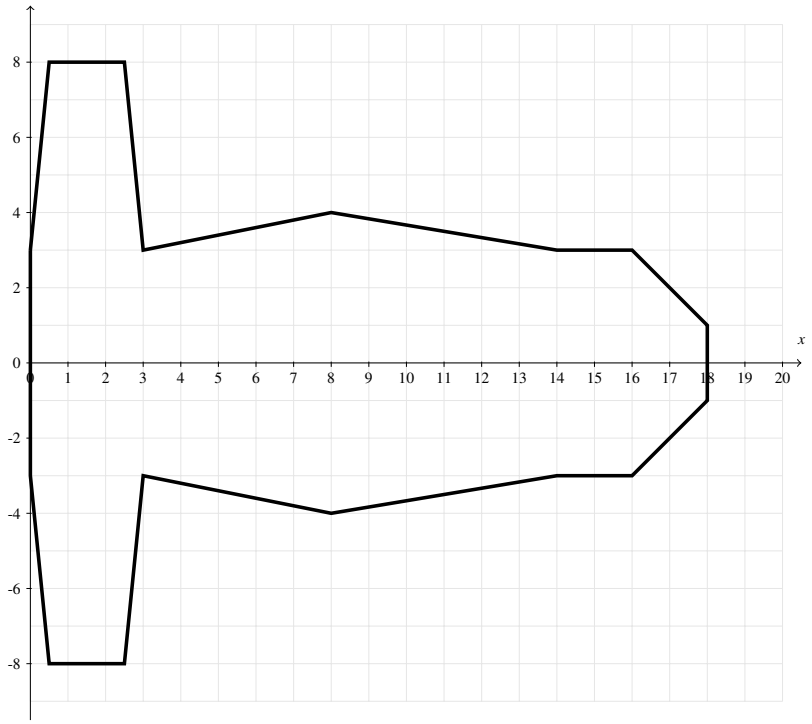
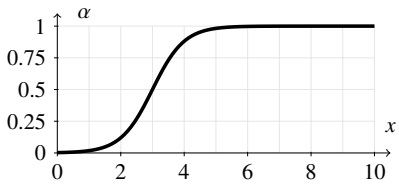
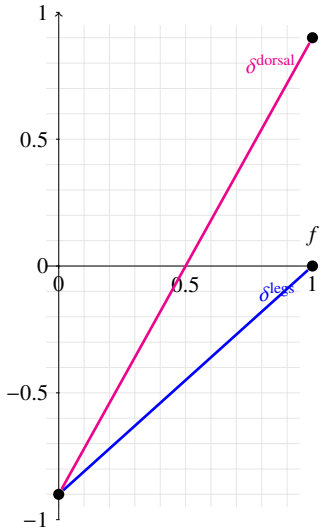


Figure 6: Spatial domain.



(a) $\kappa(x, y)$



(b) δ^{legs} and δ^{dorsal}

Top: The domain used for reaction-diffusion simulations. Left: Distribution of preparatory parameter κ as a function of transversal position. Right: Dorsal and leg potentials as a function of mixture proportion f.

Paper E

Structure identification in high-resolution transmission electron microscopy images: an example on graphene

Structure identification in high-resolution transmission electron microscopy images: an example on graphene

Jacob S. Vestergaard[a], Jens Kling[b,c], Anders B. Dahl[a], Thomas W. Hansen[b,c], Jakob B. Wagner[b], Rasmus Larsen[a]

[a]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Building 324/130, Richard Petersens Plads, 2800 Kgs Lyngby, Denmark. Phone: +4545253427.*
[b]*Center for Electron Nanoscopy, Technical University of Denmark, Fysikvej, Building 307, 2800 Kgs Lyngby, Denmark*
[c]*Center for Nanostructured Graphene (CNG), Technical University of Denmark, Ørsteds Plads 345E, 2800 Kgs Lyngby, Denmark*

Abstract

A connection between microscopic structure and macroscopic properties is expected for almost all material systems. High-resolution transmission electron microscopy is a technique offering insights into the atomic structure, but the analysis of large image series can be time-consuming. The present work describes a method to automatically estimate the atomic structure in two-dimensional materials. As an example graphene is chosen, in which the positions of the carbon atoms are reconstructed. Lattice parameters are extracted in the frequency domain and an initial atom positioning is estimated. Next, a plausible neighborhood structure is estimated. Finally, atom positions are adjusted by simulation of a Markov random field model, integrating image evidence and the strong geometric prior. A pristine sample with high regularity and a sample with an induced hole are analyzed. False discovery rate large-scale simultaneous hypothesis testing (FDR-LSSHT) is used as a statistical framework for interpretation of results. The first sample yields, as expected, a homogeneous distribution of carbon-carbon bond lengths. The second sample exhibits regions of shorter carbon-carbon bond lengths with a preferred orientation, suggesting either strain in the structure or a buckling of the graphene sheet. The precision of the method is demonstrated on simulated model structures and by its application to multiple exposures of the two graphene samples.

Keywords: graphene, structure identification, grid matching, Markov random fields, HRTEM, LSSHT

1. Introduction

Graphene has over the last 10 years received massive attention and been studied intensively due to its low mass, high strength and electrical properties (Geim & Novoselov, 2007). These properties are related to the regular two-dimensional honeycomb lattice in which carbon atoms arrange themselves in graphene. Investigations and theoretical results (Girit et al., 2009; Wang et al., 2012) predict a connection between the microscopic structure of graphene and its macroscopic properties. Therefore knowledge of the atomic structure of graphene is essential.

In this paper, a methodology for automatically determining variations in the atomic structure of a graphene sample imaged using high-resolution transmission electron microscopy (HRTEM) is provided. HRTEM makes it possible to image graphene at atomic level. However, imaging is challenging due to low mass-thickness and limited stability under the electron beam (Meyer et al., 2008, 2012). As a consequence, the images mostly show low contrast and makes it challenging to recognize the carbon atom positions precisely. The presented method aims for large-scale, low-contrast, automated structure detection in graphene. In Kling et al. (2014) we motivate the need for this method and demonstrate its use without further description of the methodology.

Manual or user guided annotations of the atomic structure have been used for various studies (e.g., Kotakoski et al., 2011; Warner et al., 2012). Clearly this is laborious and constrains the work to small-scale studies. On a larger scale, Nolen et al. (2010) segmented graphene samples into single-layer and bi-layer areas using automated image processing, but without explicitly placing the atoms. Wang et al. (2014) suggest simple morphological image operations to emphasize the grid structure, but without automatically determining the grid structure. Thus it is not possible to extract reliable atomic positions based on this method. Eder et al. (2014) use a template matching approach

to follow the transformation from graphene to glass under the electron beam and quantify structures different from hexagons. Local C-C bond length changes are not taken into account.

The methodology presented in this work aims on extracting single atom positions on a larger scale in single exposure low-contrast images of defect free hexagonal 2D structures. It is comprised of four steps: 1) Determination of global lattice properties from 2D Fourier analysis, 2) point initialization from local minima, 3) neighborhood estimation, and 4) fine adjustment of the grid using a Markov random field (MRF) formulation inspired by Hartelius & Carstensen (2003). The prior information about the lattice geometry obtained in step 1 is honored in the following three steps and incorporated in the MRF model in the final step.

We note that the simplest cases, such as a piece of pristine graphene, could easily be handled by less involved methods, e.g., local minima detection and a Delaunay triangulation. However, when small irregularities are present (e.g., due to stress, strain or imaging conditions), causing the material to differ from the expected structure, such a method does not provide a framework for handling this. If all graphene samples could be considered to be perfectly regular, there would be no need for analysis in the first place.

Since the seminal work by Geman & Geman (1984), Markov random fields have been extensively used to model spatial dependencies in image analysis. With the publicly available min-cut/max-flow algorithm by Boykov & Kolmogorov (2004) a (very fast) global solution to the simplest MRF model, the Ising model (Ising, 1925), could be obtained. Common for all of these MRF models are, that a priori knowledge of the neighborhood is needed. In fact, the presence of a known neighborhood is usually why a given problem is modeled as an MRF. This knowledge is not available in advance for the problem of graphene structure identification. The neighborhood needs to be estimated *before* posing the grid alignment problem as an instance of a Markov random field. This adds to the difficulty of the problem and is therefore part of this work.

The methodology is described in detail and applied to model cases and two experimental image series. A statistical interpretation of the estimated structures from simulation studies and the two cases are given. Specifically, the carbon-carbon bond lengths are treated under the principle of large-scale simultaneous hypothesis testing (Efron, 2004).

2. Data

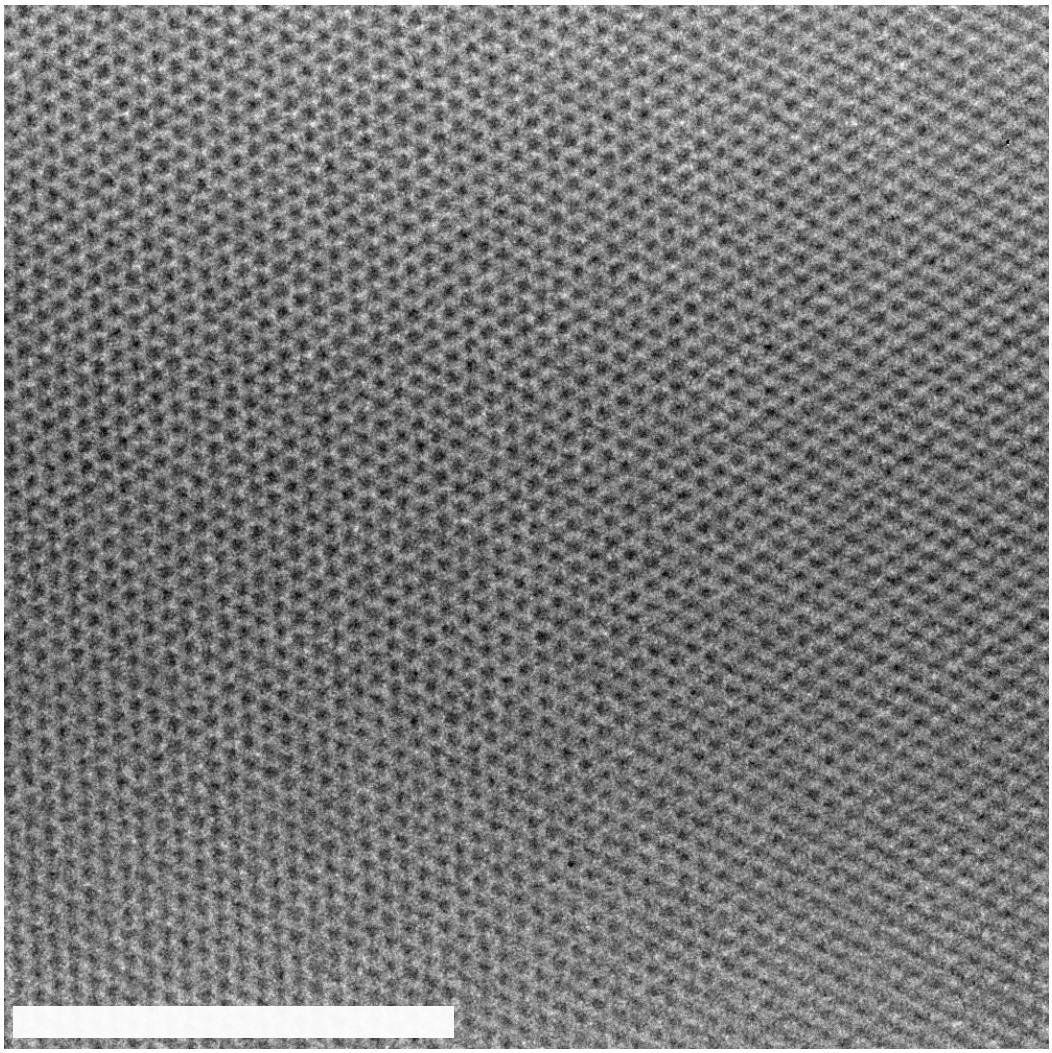
In order to obtain the experimental data we imaged suspended single-layer graphene samples with a FEI Titan environmental transmission electron microscope (ETEM). Samples are produced by either chemical vapor deposition (Graphenea, San Sebastian, Spain) or exfoliated from graphite (Booth et al., 2008) and transferred to TEM grids. The microscope is equipped with a monochromator at the gun and a spherical aberration (C_s)–corrector for the objective lens. All images are acquired with the microscope operated at 80kV and recorded on a US1000 CCD (Gatan, Pleasanton, USA) with an exposure time of 1s. A resolution better than 1.2Å was obtained by optimizing the imaging conditions.

HRTEM images are simulated using the software JEMS (Stadelmann, 2004). Multislice parameters were chosen according to the used FEI Titan with an energy spread of 0.3eV, negative Cs and positive defocus. As noise the preset "uniform noise" of the software was used with noise settings from 0 to 5%.

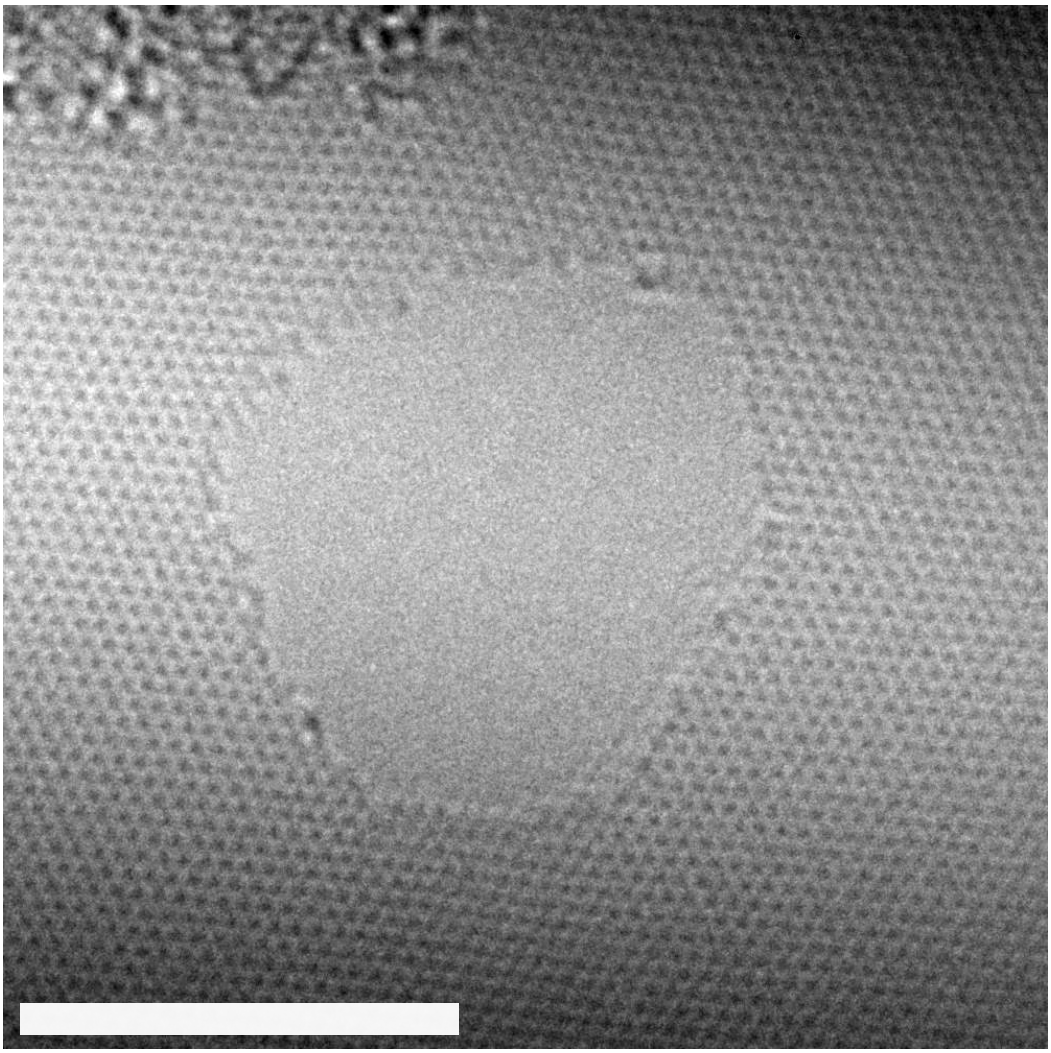
Two cases are considered. The first is a region of pristine graphene (Figure 1a) and the second is a region with an induced hole in the graphene, formed under the influence of the electron beam (Figure 1b). Pristine graphene is naturally the simplest case possible for any automated method and is included to have a baseline comparison for the second case. The second case is much more challenging due to the irregularities arising around this alteration of the structure. This will become apparent in Section 4.

Notice in Figure 1 that the graphene structure is nicely resolved, even though individual carbon atoms are not obvious. Due to the present imaging conditions, meaning negative Cs and positive defocus, the carbon hexagons appear bright, where the centers of the hexagons are dark. Knowledge of these microscope parameters are crucial, as positive Cs or different defocus/z-height can lead to an inverse contrast in the image. This is important when considering the lattice geometry next.

The only manual interaction during the process was discarding the area in the upper-left corner of the altered graphene sample, where amorphous material is present. An automated segmentation step could easily have handled this, but is not implemented here. Note that the hole in the middle was not manually discarded, but left to the image processing pipeline to deal with.



(a) Pristine graphene



(b) Altered graphene

Figure 1: Two different graphene samples used as case studies throughout this work. a) Excerpt of a piece of pristine graphene. b) Excerpt of a graphene sample with an induced hole-defect. The scale bar is 5 nm.

3. Methodology

When estimating a complete lattice structure consisting of say 5000 separate atom positions, robustness is of the essence. An image processing pipeline is created, progressing from global to local processing, carrying estimates of global properties forward and letting local mechanisms handle fine adjustment. This ensures robustness while also keeping the computational burden low.

The four steps of the image processing pipeline are presented below.

Parameters describing the global properties of the lattice are estimated first. The most important property estimated is the scale. Next, a first guess at the hexagon center positions is obtained by classical image analysis techniques, i.e. contrast enhancement and local extrema detection. Third, a plausible neighborhood structure is created. This lays the ground for the fourth and final step, namely refining the positioning of the hexagon centers, while respecting both the image evidence and the geometric prior.

In the first step, the lattice geometry in connection to its 2D Fourier representation will be presented.

3.1. Lattice parameters from Fourier analysis

We consider the observed image $\mathcal{I} \in \mathbb{R}^{m \times m}$. It is assumed that this image primarily consists of a regular hexagonal lattice with carbon-carbon (C-C) bond length as hexagonal side length t.

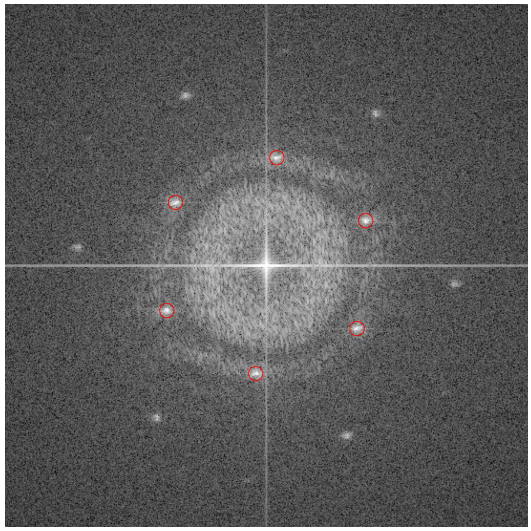


Figure 2: Excerpt of a 2D Fourier log-magnitude image. The red circles correspond to detected peaks from the first reflection with distances $r_i, i = 1, \ldots, 6$.

2D Fourier analysis of the image yields a log-magnitude image, such as illustrated in Figure 2. The usefulness of the 2D Fourier analysis in this context is well known, see for instance Meyer et al. (2008) or Zhang et al. (2009). The spots represent the reciprocal lattice planes belonging to the {100} and {110} lattice planes of graphene with the distance r for the {100}/{1-10} reflections.

The spots are detected in the Fourier magnitude image by local maxima detection and the points of detection are represented in polar coordinates as $p_i = (\phi_i, r_i), i = \{1, \ldots, 6\}$. The smallest angle relative to horizontal is directly translated into the rotation ϕ of the lattice. The average radius $\bar{r} = \frac{1}{6} \sum_{i=1}^{6} r_i$ is transformed from Fourier space to pixel distances and, by geometric relations, used to determine the hexagonal side length in pixels t:

$$t = \frac{2m}{3\bar{r}} \tag{1}$$

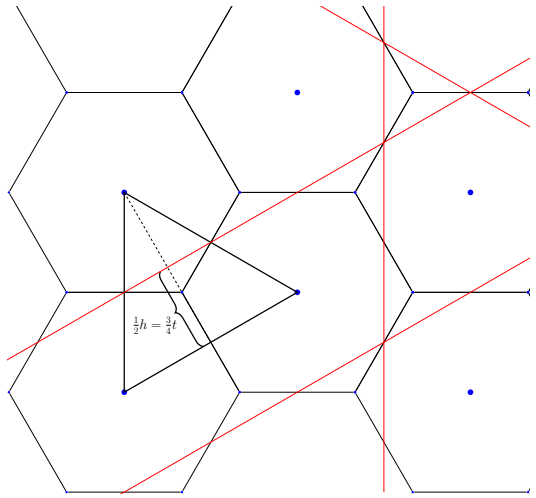


Figure 3: Hexagonal mesh with red lines representing the {100} lattice planes with distance h. The drawn triangle in black with height $h = \frac{3}{2}t$ and side length $\sqrt{3}t$ illustrates the geometrical relationship between the hexagonal tessellation and its dual, the triangular.

where m is the width of the image in pixels.

The estimated hexagonal side length is used as a scale parameter to leverage knowledge about the lattice geometry in the next steps.

3.2. Hexagon center detection

The primary (hexagonal) grid is not always evident from the image, wherefore a dual (triangular) lattice reconstruction is aimed for. The triangular lattice can be seen as a graph with vertices being the hexagon centers and edges connections between these.

First, standard blob enhancement techniques (Lindeberg, 1996) are used for enhancing circular areas with a scale of $\frac{1}{2}t$. Next, the image contrast is enhanced using the top-hat contrast operator (Soille, 2003). A disc with radius $\frac{1}{2}t$ is used as structuring element. After the image referred to as $\mathcal{I}$ has undergone these operations, it will be referred to as $\mathbf{Y}$. Finally, local minima are detected in $\mathbf{Y}$. These minima serve as a first guess on hexagon center positions.

3.3. Neighborhood

The problem of forming a meaningful graph from the candidate hexagon centers $C = \{c_i, i = 1, \ldots, N\}$ is considered. "A meaningful graph" is a graph where neighboring nodes are centers of actual neighboring hexagons. This neighborhood is needed to eventually construct the dual lattice (the actual atom positions), but will also be necessary for the fine adjustment step presented below (Section 3.4). The method described below takes spatial consistency into account to avoid spurious holes in the mesh that might otherwise occur due to low contrast in the image.

Two elements are taken into consideration, namely the image intensities under each center point and the regularity of the triangles. Both are instances of a problem where the range of each observation x is an interval $[a_1, a_2]$ with one of the endpoints being the optimal value. When observing intensities at the hexagon centers, the range will be $x_c \in [0, 1]$ with an optimal value of $x_c = 0$ in the image $\mathbf{Y}$ equivalent to black. When considering a triangle's regularity, the observed term is the triangle's minimum angle in the range $x_\triangle \in [0, \frac{\pi}{3}]$. This should ideally be $x_\triangle = \frac{\pi}{3}$, i.e., the observed triangle is equilateral. Note that subscript c is referring to intensities under the hexagon centers and subscript $\triangle$ to the minimum angle of a triangle.

A simple, yet effective, scheme achieving this is presented here.

The Delaunay triangulation is of great aid as it maximizes the minimum angle of all triangles in the triangulation. However, the Delaunay triangulation cannot be used directly, since not all triangles are physically meaningful. To overcome this an iterative scheme is implemented, consisting of the following steps:

1. Delaunay triangulation of C
2. Remove improbable triangles
3. Remove improbable and unreferenced points from C
4. If points were removed, goto step 1

To remove the improbable triangles and points, an Ising model is formulated as a minimum cut problem (Boykov & Kolmogorov, 2004). An illustration of the model can be seen in Figure 4. A tractable property of this model is spatial smoothness and the possibility to obtain the optimal solution using the min-cut/max-flow algorithm.

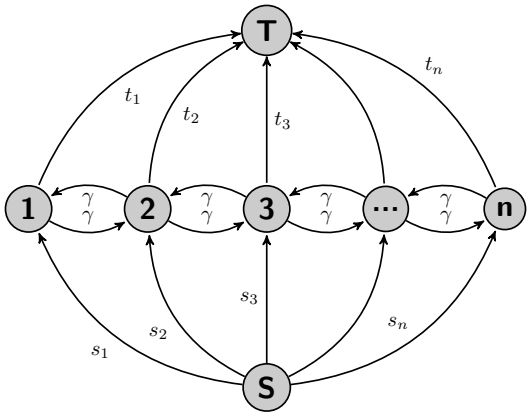


Figure 4: Graph representation of the Ising model. The numbered entities – the sites – are triangles and pixels for each of the two steps in the iterative scheme.

The terminal weights s_i and t_i are formulated from a simple convex combination of the Euclidean distance from the observed value x_i to the two interval end points a_1, a_2.

$$s_i = (1 - \alpha)\|x_i - a_1\|_2$$
$$t_i = \alpha \|x_i - a_2\|_2 \ .$$

The regularization term α controls the tolerance of the distance and we use a constant neighborhood term γ. Notice that for $\alpha = 0.5$ this is simply assigning values closest to a_1 to the sink and values closest to a_2 to the source. This way of assigning terminal weights makes no assumptions on the distribution of the observed values.

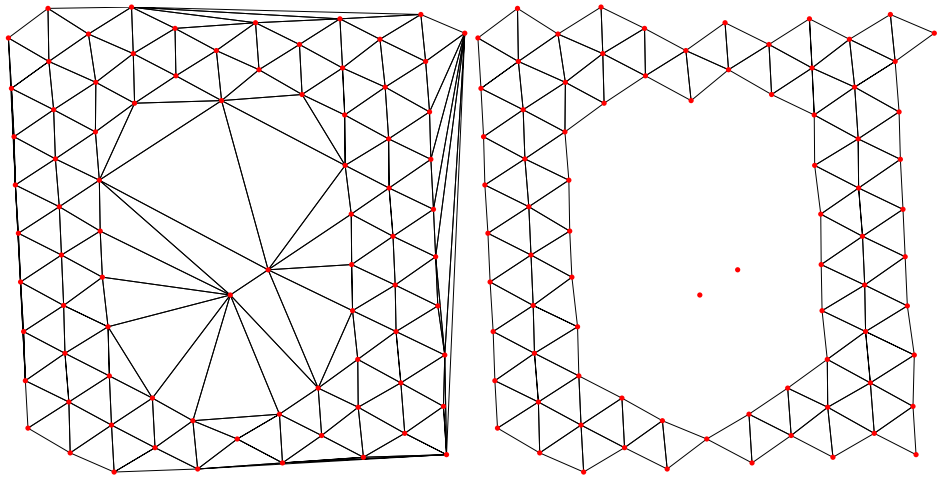
Values of $\gamma_\triangle = 3$ and $\gamma_c = 2$ were found to be a good choice for neighborhood terms and $\alpha = 0.85$ a suitable regularization parameter for both models.

An example of the procedure can be seen in Figure 5, where a triangulation is cleaned up to only represent meaningful connections between nodes. For this simple example, only a single iteration was needed.

3.4. Fine adjustment by Bayesian grid matching

The iterative scheme presented above renders a meaningful neighborhood, but the hexagon centers are still placed where the initialization in Section 3.2 placed them. These positions are good, but not optimal in the sense that they do not consider the prior geometric knowledge. An undirected graphical model (an MRF) that takes this into account is presented here.

Inspired by the Bayesian formulation of the grid matching problem in Hartelius & Carstensen (2003) a set of node-sites $\mathcal{S} = \{s_i, i = 1, \ldots, n\}$ in a graph is observed. The set $\mathcal{L} = \{l_{ij}, i \sim j, i \leq j\}$ represents the arcs in the graph, where l_{ij} connects node-sites s_i and s_j, also denoted by $i \sim j$. The number of neighbors for node s_i is denoted as n_i. Locations of the grid nodes are contained in the list $\mathcal{G} = \{\mathbf{g}_i, i = 1, \ldots, n\}$ where $\mathbf{g}_i = [x_i, y_i]^T$ is the location of node s_i.



(a) (b)

Figure 5: Example of the iterative procedure used to construct the initial neighborhood.

A neighborhood system for both nodes and arcs is defined in Figure 6. The node neighborhood is defined as the one-clique in the triangular lattice. An arc's neighborhood includes other arcs from the same simplices. $ij \sim kl$ denotes that arc l_{ij} and l_{kl} are neighbors and n_{ij} denotes the number of arc neighbors for l_{ij}.

Coding scheme. In theory each node should be visited in random order. In practice a coding scheme is employed, such that multiple nodes are visited simultaneously. To obey the Markov property, the coding scheme is constructed such that two neighboring nodes are not altered simultaneously. Since the lattice is deduced from the local minimum, the coding scheme cannot be explicitly constructed. Rather it is an instance of the graph coloring problem to which a (sub-optimal) solution is obtained using the Welsh-Powell algorithm (Kubale, 2004; Welsh & Powell, 1967).

Geometric prior. Two suggestions for simple geometric priors are outlined here. The first is to model the expected arc length as constant throughout the sample, i.e., the distance between two neighboring nodes is modeled as

$$\|\mathbf{g}_i - \mathbf{g}_j\|_2 = \bar{r} + \epsilon_{ij} \tag{2}$$

where $\epsilon_{ij} \in N(0, \sigma^2)$.

The second model is more flexible, where the distance between two neighboring nodes is modeled as

$$\|\mathbf{g}_i - \mathbf{g}_j\|_2 = \bar{r} + t_{ij} + \epsilon_{ij} \tag{3}$$

where $\epsilon_{ij} \in N(0, \sigma^2)$ and

$$t_{ij} = \frac{1}{n_{ij}} \sum_{ij \sim kl} \|\mathbf{g}_k - \mathbf{g}_l\| - \bar{r} \tag{4}$$

is the average deviation from the expected arc length $\bar{r}$ in the neighborhood of arc l_{ij}. This allows for a smooth change from the expected arc length over the lattice, while introducing a small anistropy favoring deviations parallel to the three C-C bond orientations due to the chosen arc neighborhood. In a perfect triangular lattice $t_{ij} = 0$. Throughout this paper, the locally adaptive geometric prior in Equation (3) has been employed.

The joint distribution of $\mathcal{G}$ is given by

$$P(\mathcal{G}) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{l_{ij}} \left(\|\mathbf{g}_i - \mathbf{g}_j\|_2 - (\bar{r} + t_{ij})\right)^2\right]. \tag{5}$$

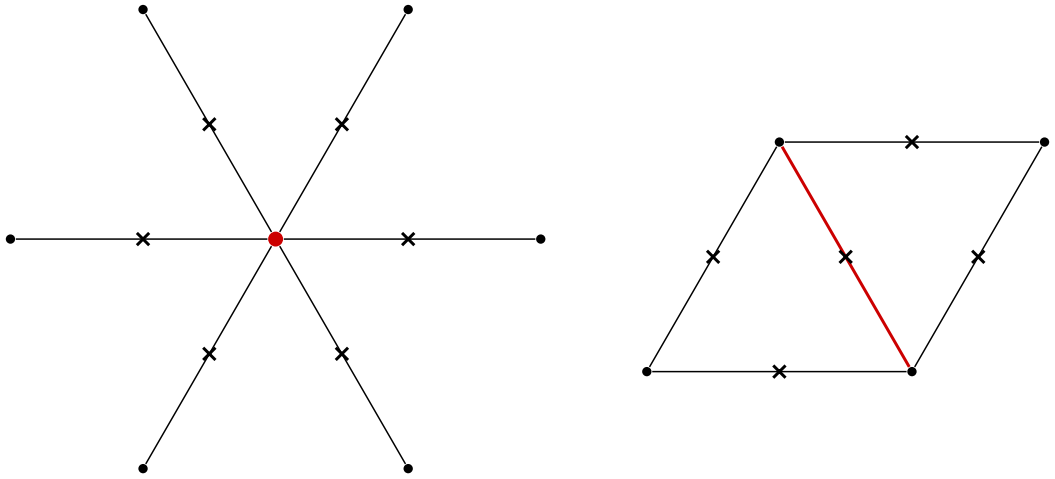


Figure 6: Each node (marked in red) has six node neighbors (dots) and thus six arc neighbors (crosses). The arc (marked in red) has four arc neighbors (crosses).

This is modeled as a Markov Random Field (MRF) and thus the probability of node position $\mathbf{g}_i$ is only dependent on its neighbors

$$P(\mathbf{g}_i|\{g_j, j \neq i\}) = P(\mathbf{g}_i|\mathbf{g}_j, i \sim j)$$
$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i\sim j}\left(\|\mathbf{g}_i - \mathbf{g}_j\|_2 - (\bar{r} + t_{ij})\right)^2\right] . \quad (6)$$

The Gibb's representation of this geometric prior is $P(\mathcal{G}) \propto \exp\{-U(G)\}$.

Aposteriori lattice estimation. The pre-processed image $\mathbf{Y}$, introduced previously, is considered as the observed image. The observation model $P(\mathbf{Y}|\mathcal{G})$ describes the probability of observing the image $\mathbf{Y}$ given the configuration $\mathcal{G}$, i.e., the likelihood. For a single site s_i

$$P(\mathbf{Y}|\mathbf{g}_i) \propto \exp\{-\mathbf{Y}(x_i, y_i)\} = \exp\{-U(\mathbf{Y}|\mathbf{g}_i)\} .$$

The likelihood energy $U(\mathbf{Y}|\mathbf{g}_i)$ is simply the intensity value of $\mathbf{Y}$ at $\mathbf{g}_i$, thus a darker spot in $\mathbf{Y}$ corresponds to a lower energy than a bright spot.

The posterior distribution of the model can then be written as

$$P(\mathcal{G}|\mathbf{Y}) = P(\mathcal{G})P(\mathbf{Y}|\mathcal{G}) \quad (7)$$

where the observation model represents the faith in the observed data and the geometric prior enforces regularity in the lattice.

Simulated annealing. The Gibb's representation of the posterior distribution is

$$P(\mathcal{G}|\mathbf{Y}) \propto \frac{1}{Z}\exp\left\{-\frac{U(\mathcal{G}) + U(\mathbf{Y}|\mathcal{G})}{T}\right\}$$

where Z is a normalizing constant and T the temperature of the system.

A simulated annealing scheme is employed to maximize Equation (7). The Metropolis spin-flip algorithm is used to generate moves in the random walk, similar to Hartelius & Carstensen (2003):

1. Start with configuration $\mathcal{G}$.
2. Choose a node s_i and take a step to obtain $\mathbf{g}_i'$.
3. Set configuration $\mathcal{G}'$ equal to $\mathcal{G}$ with node s_i's position set to $\mathbf{g}_i'$.
4. Replace $\mathcal{G}$ with $\mathcal{G}'$ with probabilty

$$\begin{aligned} p &= \min(1, P(\mathcal{G}')/P(\mathcal{G})) \\ &= \min(1, \exp(-U(G') + U(G)) \end{aligned} \tag{8}$$

5. if not stop, go to step 2.

The relation in (8) is of practical importance, since this allows us to work on energy differences rather than ratios of probabilities.

Generating $\mathbf{g}_i'$ in step 2 can be done in numerous ways. Here $\mathbf{g}_i$ is chosen as one of the eight neighboring pixels.

The temperature scheme for the simulated annealing is chosen to start from $T_0 = 0.1$ and decrease to $T_{\text{end}} = 10^{-10}$ over $n_{\text{iter}} = 1000$ iterations. The temperature at iteration k is assigned according to $T_k = c \cdot T_{k-1}$, where c is determined from the number of iterations and temperature end points. During each iteration, each node site is visited four times before decreasing the temperature.

Atom placement. The dual lattice to the triangular forms the hexagonal lattice of C atoms. Therefore the atoms are placed in the centers of the fitted triangles. Thus three hexagonal centers are used to position each atom.

A movie visualizing the optimization while running is included as Supplementary movie1.

4. Results and statistical interpretation

The two samples of graphene (one pristine and one altered) are processed using the described image processing pipeline. The result of this pipeline is a set of hexagon center positions and a neighborhood, i.e., a graph. This graph is used to derive the atom positions (the dual lattice) and their mutual distances, the carbon-carbon bond lengths. These lengths are the primary derived measure and will be the subject of the analyses in this section. Such a result is available at two stages of the pipeline, namely before and after fine adjusting the positions.

First, the effect of this fine adjustment will be quantified by simple statistics. Following this, the final positioning will be analyzed using a statistically sound visualization allowing for objective interpretation.

Figure 7 summarizes graphically the effect of fine adjusting the hexagon center positions. The carbon-carbon bond lengths are illustrated as box plots, summarizing simple statistics. A total of 5057 and 3589 atoms were positioned with 14966 and 10452 bonds estimated respectively. The two left boxes concern the pristine graphene case and the two right boxes concern the altered graphene case. The first box for each case summarizes the C-C bond lengths if one had accepted the positions determined by the initial grid procedure in Section 3.2. The second box shows the same statistics, but for the final result, i.e., after the hexagon centers have been fine adjusted using the procedure in Section 3.4.

First off, an inter-case comparison suggests that much higher variability is present in the sample, where the structure has been altered. This says more about the physical properties of the samples, than the properties of the fine adjustment procedure.

More interesting is it to consider the intra-case variability. It is evident that this has been greatly reduced in both cases by the fine adjustment procedure. This can also be visually verified by inspecting a coloration of the C-C bonds in Supplementary material 2. Numerically, the standard deviation has been reduced from $4.39 \cdot 10^{-3}$nm to $2.11 \cdot 10^{-3}$nm and $7.31 \cdot 10^{-3}$nm to $3.31 \cdot 10^{-3}$nm respectively. Both reductions are significant changes of variance with p-values $\ll 0.01$ according to a Bartlett's F-test with one degree of freedom.

In order to interpret the estimated atomic structure in a statistical setting, we use the framework of false discovery rate controlled large-scale simultaneous hypothesis testing (FDR-LSSHT) proposed in Efron (2004). Specifically, the structure is evaluated in terms of the distribution of carbon-carbon bond lengths.

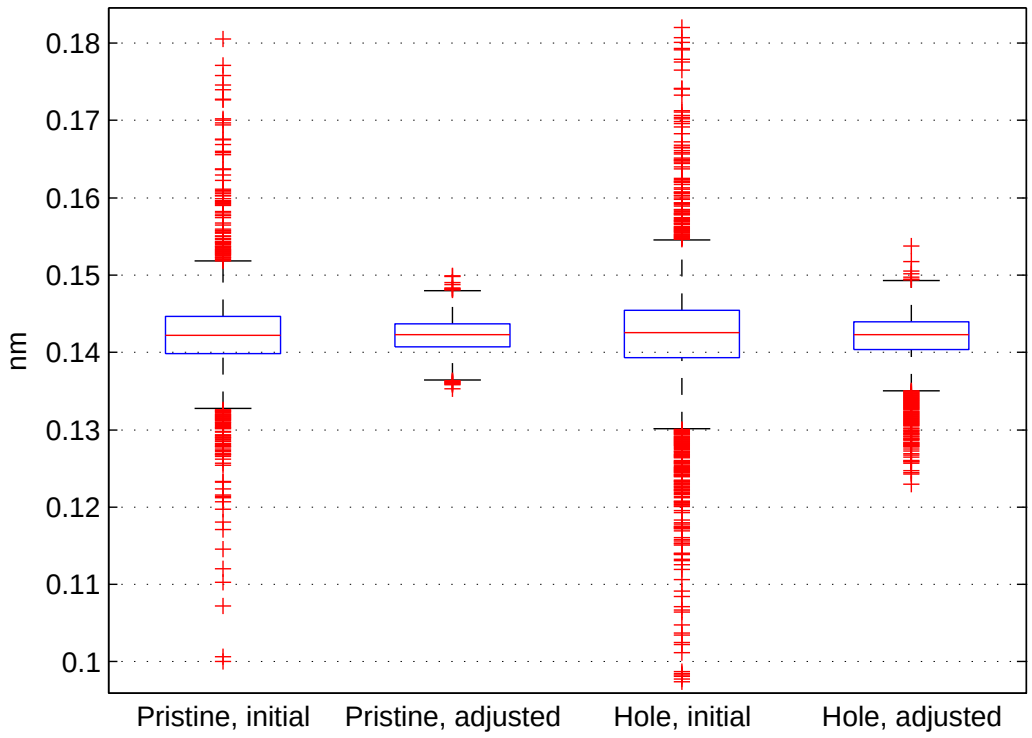


Figure 7: Box plots summarizing the distribution of estimated C-C bond lengths before and after fine adjustment of the hexagon centers. The edges of the box gives the first and third quartile of the distribution, the horizontal line within the box represents the median. The whiskers are 1.5 times the interquartile range of length.

4.1. Large-scale simultaneous hypothesis testing

In statistical testing of small sample-large number of variables cases normal multiple comparison procedures (e.g. Bonferroni adjustment) tend to be too conservative and leading to too few significant variables. FDR-LSSHT is a framework for simultaneous evaluation of a large number of hypothesis tests while controlling the significance testing by setting a maximum for the proportion of false positives.

In this setting we work on z-values, where

$$z_i = \frac{x_i - \mu}{\sigma} \ , \quad i \in \{1, \ldots, N\}.$$

Here, x_i is the length of the i'th carbon-carbon bond and (μ, σ) are parameters for a normal distribution under the null hypothesis. Histograms of x_i for the two cases are shown in Figure 8.

Based on these z-values, FDR-LSSHT is a three-step process:

1. estimate the distribution of z-values $f(z)$ as a smooth spline fitted to the sample histogram (green in Fig. 8),
2. find an empirical null hypothesis $f_0(z)$ (magenta in Fig. 8) and
3. calculate the false discovery rate $fdr(z) = f_0(z)/f(z)$.

"Interesting" observations will then show up with a low false discovery rate (Efron suggests $fdr(z) < 0.1$). Here, "interesting" means observations not following the dominant normal distribution $f_0(z)$.

A cubic spline with twenty evenly distributed knots is used, which is least-squares fitted to the square-root of the sample histogram counts to represent $f(z)$.

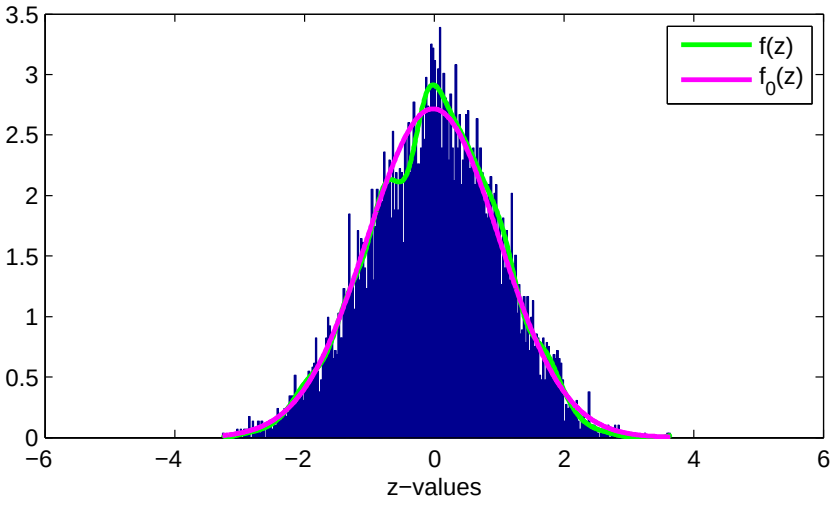
The empirical null hypothesis $f_0(z)$ is estimated as a Gaussian $\mathcal{N}(\mu_0, \sigma_0^2)$ considering the width of the distribution at half its maximum. However, care should be taken; inspection of the histogram in Figure 8b clearly shows a heavier tail to the left. Therefore, the distance from the peak's position μ_0 to the z-value z_{half} fulfilling $f(z_{\text{half}}) = \frac{1}{2} f(\mu_0), z_{\text{half}} > \mu_0$ is used to obtain

$$\sigma_0 = \frac{z_{\text{half}} - \mu_0}{1.17741} \ .$$

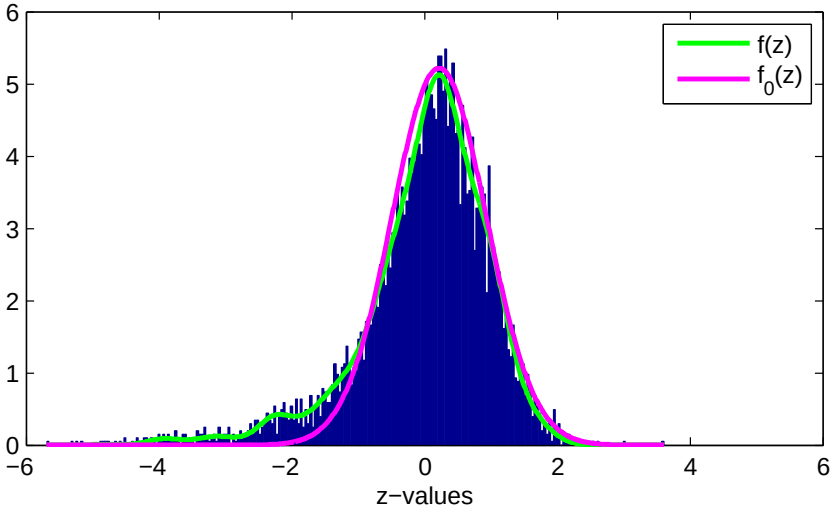
For further considerations on estimation of $f_0(z)$ the reader is referred to Efron (2004).

4.2. False discovery rate

The false discovery rates are overlaid their respective C-C bonds in Figure 9 according to a color scale. Observations with an FDR above 0.2 are colored white, as they are considered not interesting. This scale is chosen based on Efron's recommendation of considering FDRs below 0.1 interesting. Notice how this statistically sound visualization translates nicely into a visualization of untypical areas in the graphene. For the pristine graphene in Figure 9a, a homogeneous distribution is observed, as expected. Only few C-C bonds at the border of the image stand out and are most likely related to artefacts.

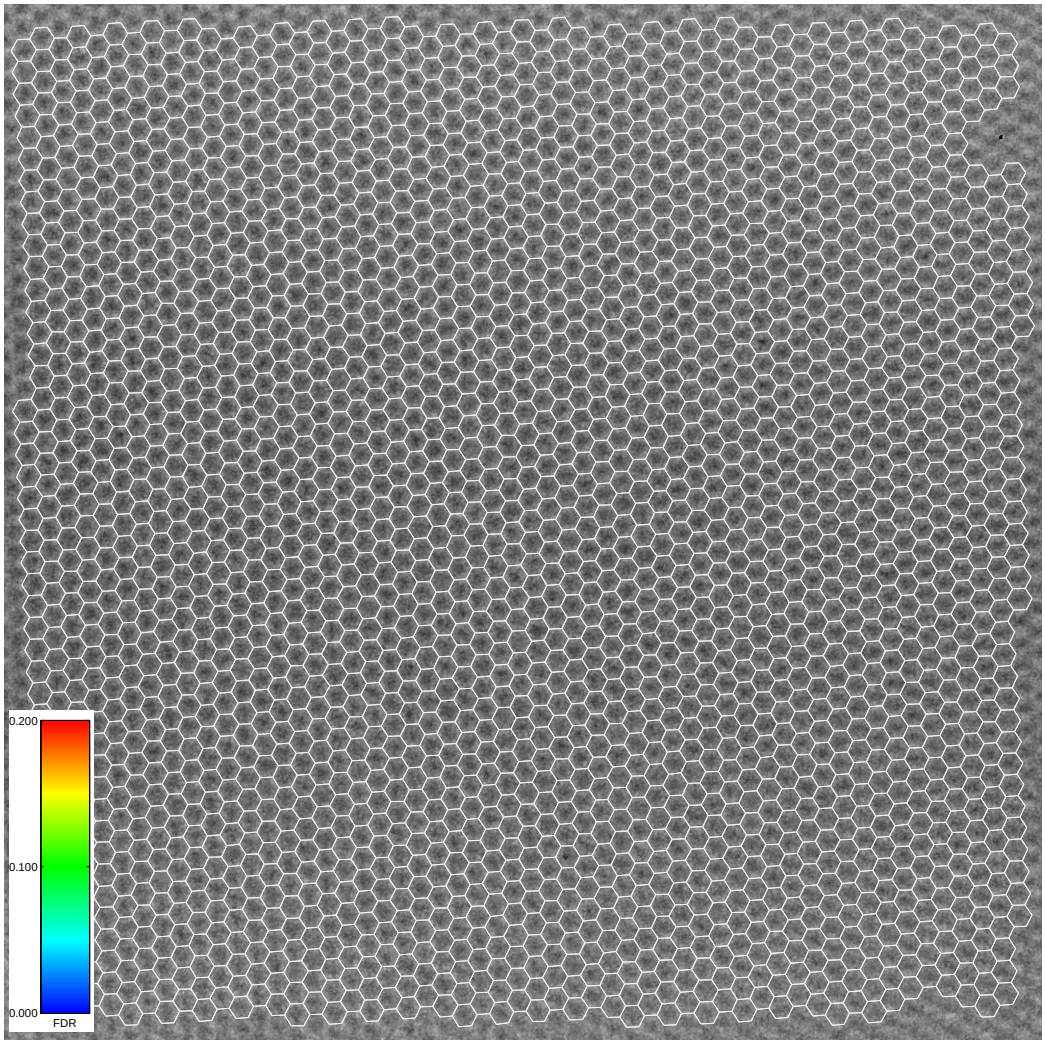


(a) Pristine graphene

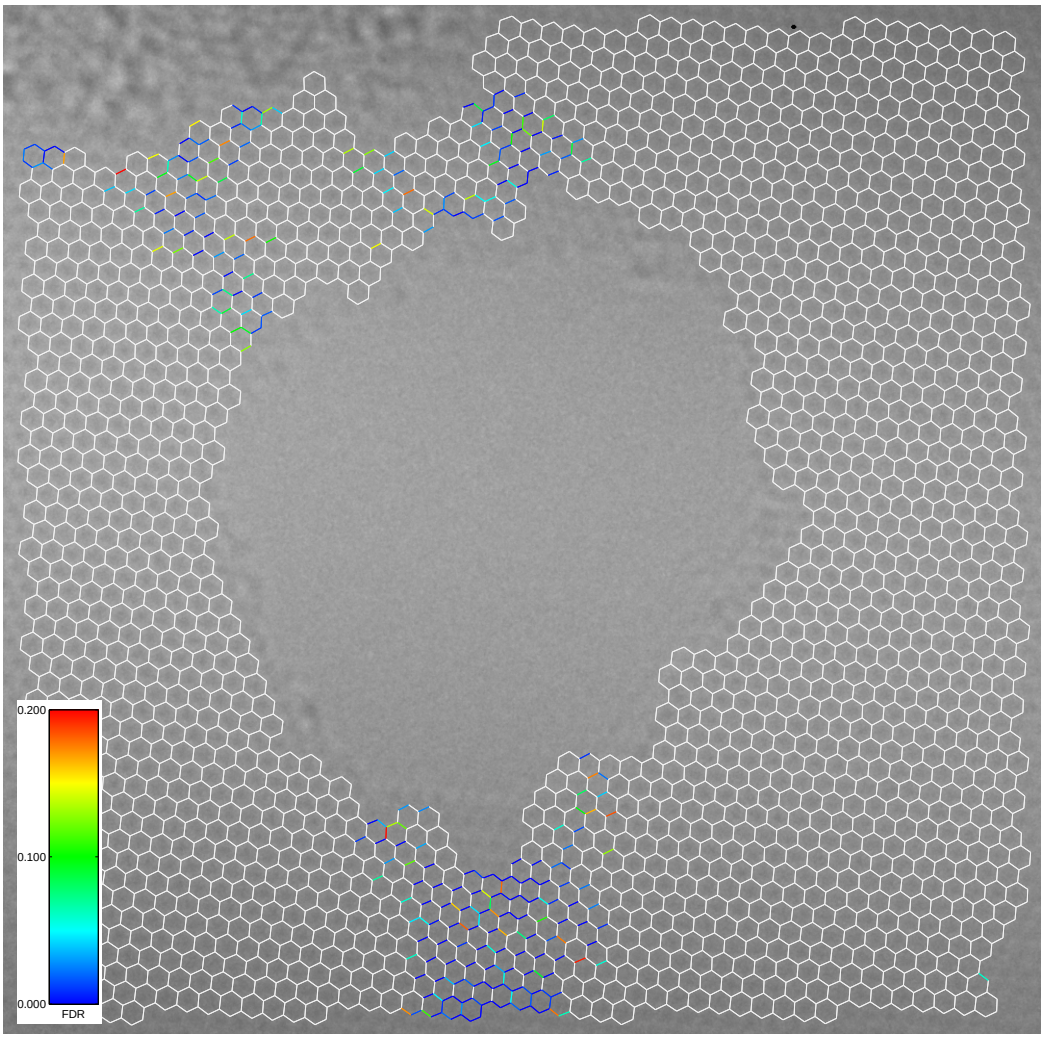


(b) Graphene with induced hole

Figure 8: The empirical null hypothesis $f_0(z)$ and estimated distribution of z-values $f(z)$ for both cases. Notice the heavy lower tail on $f(z)$ in (b) compared to the empirical null hypothesis.



(a) Pristine graphene



(b) Graphene with induced hole

Figure 9: C-C bonds colored according to their false discovery rate. The color scale ranges from 0.0 to 0.2 where a low false discovery rate (blue) indicates "interesting" observations, i.e., observations not following the empirical null hypothesis. Notice how areas above and below the induced hole-defect are highlighted, while the pristine case shows no such pattern.

The pristine areas around the induced hole in the second case reveal the same homogeneous distribution as in the first case.

Most of the bonds having a low FDR are also short in length, as seen in the histogram before. These shorter bonds reveal a preferred orientation, approximately in horizontal direction. Physically this could either show strain in the graphene structure, or represent a 3D information. Suspended graphene is known to form out of plane ripples (Meyer et al., 2007). Bond lengths appear shorter as the structure is buckled or folded, as only a projected bond length is visible. Together with the preferred orientation, this would speak for a folding. Neither can these two possibilities be discriminated, nor a combination of both excluded. Further investigations are ongoing.

4.3. Precision

To verify the reliability and precision of the method, HRTEM images of graphene model structures are simulated (see Supplementary material 1). The structures vary in bond length, from ideal graphene to a strongly strained C-C bond length. This does not have to be physical, but gives an indication of the reliability and precision of the method. Furthermore, different noise levels are added to the images to resample less contrast in a real experiment. The algorithm recognizes the structures precisely and extracts the C-C bond lengths. This even works for rather high noise levels. Specifically the simulation studies show that C-C bond lengths can be estimated with a bias in the range of 0.0001nm–0.0004nm depending on the simulated strain. A precision of less than 0.0010nm is achieved in noise free cases and up to 0.0024nm for simulations with high noise.

Additionally, multiple exposures of each graphene sample are used. Ten exposures were available of the pristine graphene sample and twenty exposures of the sample with an induced hole. The same region of interest and parameters as above were used for all exposures.

The carbon-carbon bond lengths are extracted for each exposure. A spline with twenty knots is fitted to the bond length histogram on the domain $[0.122, 0.162]$nm, similar to the FDR-LSSHT procedure above. Each spline represents a distribution of carbon-carbon bond lengths. All of these distributions are shown in Figure 10. Blue lines represent exposures of the sample with an induced hole. Red lines represent exposures of the pristine graphene sample.

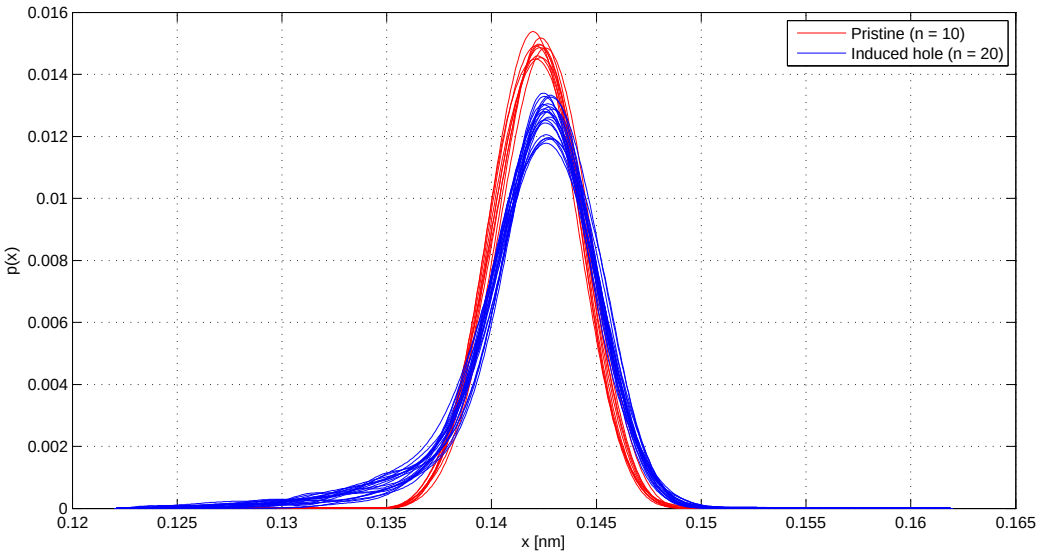


Figure 10: Splines with twenty knots fitted to the carbon-carbon bond length histograms for multiple exposures. Blue lines show distributions estimated for twenty exposures of the graphene sample with an induced hole. Red lines show the same for ten exposures of the pristine graphene sample.

The bond length distributions for all exposures of the area with an induced hole have the characteristic heavy tail of shorter bond lengths. This heavy tail is not present in the pristine sample.

Comparing the single exposures, the pristine case does not show a strong variation. This can be seen in the color maps of the single exposures. The structures appear homogenous (see Supplementary material 2). For the sample with the induced hole, the distributions look similar, but in the color maps a clear change can be seen. The orientation of the shorter bonds remain the same, but the area of appearance slightly moves with exposures. Nevertheless, they remain situated above and below the hole.

This observation can be correlated with an beam induced effect. Even though an accelerating voltage of 80kV is low enough to preserve the pristine graphene structure, atoms at defects and edges possess a lower threshold to be removed or rearranged (Kotakoski et al., 2012). This goes together with a deformation of the lattice. Furthermore, a chemical etching at the edge can result in a rearrangement of the atoms and a deformation (Meyer et al., 2012). A run test shows that no evidence can be found for a temporal trend in the average bond length (see Supplementary material 2 for details). The simulated images and the time sequences demonstrate that the method is sufficiently precise.

5. Discussion

The presented method is only applicable to materials forming a hexagonal lattice. However, it is possible to replace one or more of the building blocks constituting the pipeline and make it useful for other geometries, e.g., by replacing the Delaunay triangulation with something meaningful for the geometry at hand. The result for three dimensional materials orientated in a zone axis would then not be single atom positions, but positions of the atom columns along the viewing direction. We believe the approach of a rough initialization and subsequent fine adjusting of the structure a general one.

We stress that this method is currently not capable of detecting defects, e.g., penta- and heptagonal carbon rings. To incorporate this, the Markov random field formulation used for fine adjusting the grid would need to be extended, e.g., by marginalizing out the different local geometries and their respective neighbor distances. Another possibility would be to fit the best possible hexagonal lattice and in a subsequent step suggest new configurations, i.e., swap a perfect part of the lattice for a defect, and choose the one with maximum likelihood. These extensions are left for future work, but the locality of this model looks promising for making it possible.

6. Conclusions

A method to determine atomistic properties of periodic systems with low contrast in general has been presented. Graphene is used as an example of such a material. The atomic structure and atomistic parameters of large-scale graphene samples have been automatically estimated from high-resolution transmission electron microscopy images.

A pipeline consisting of four main steps for estimating carbon atom positions have been described: 1) Determination of global lattice properties from 2D Fourier analysis, 2) point initialization from local minima, 3) neighborhood estimation, and 4) fine adjustment of the grid taking prior assumptions and observed data into account. Parameters for the geometric priors are estimated in the first step and carried forward through the pipeline. It is shown that adjusting the atom positions according to a Markov random field model significantly reduces the variation of carbon-carbon bond length estimates.

Two distinct cases have been chosen to demonstrate the method's capabilities, namely a piece of pristine graphene and a piece of graphene with an induced hole. The framework of false discovery rate large-scale simultaneous hypothesis testing (FDR-LSSHT) was employed to provide a statistically sound interpretation of the resulting estimates. Specifically, the distribution of carbon-carbon bond lengths was analyzed.

It was found that the pristine graphene showed high regularity in the atom positions and thus the distribution of C-C bond lengths were approximately normal distributed around a bond length of 0.142 nm. The graphene sample with the induced hole was found to contain areas of significantly shorter bond lengths, manifesting themselves as a heavy lower tail in the distribution of C-C bond lengths. Investigations into whether this is due to buckling of the graphene sheet or stress in the configuration are still ongoing.

The precision of the proposed method was verified on simulated images and evaluated on real data by its application to multiple exposures of each sample. Ten exposures of the pristine graphene and twenty exposures of the sample

with an induced hole were analyzed. A single set of parameters was used for all 30 exposures and the within-sample distributions were found to be very similar, which demonstrates the robustness of the method.

Future work includes applying the method in a high-throughput setting to test specific material hypotheses, such as structural changes under external stimuli like temperature or current, and relate them to observed physical measurements.

References

Booth, T. J., Blake, P., Nair, R. R., Jiang, D., Hill, E. W., Bangert, U., Bleloch, A., Gass, M., Novoselov, K. S., Katsnelson, M. I., & Geim, a. K. (2008). Macroscopic graphene membranes and their extraordinary stiffness. *Nano Letters*, *8*, 2442–6.

Boykov, Y., & Kolmogorov, V. (2004). An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*, 1124–1137.

Eder, F. R., Kotakoski, J., Kaiser, U., & Meyer, J. C. (2014). A journey from order to disorder - atom by atom transformation from graphene to a 2D carbon glass. *Scientific Reports*, *4*, 4060. doi:10.1038/srep04060.

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, *99*, 96–104.

Geim, a. K., & Novoselov, K. S. (2007). The rise of graphene. *Nature Materials*, *6*, 183–91.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–41.

Girit, C. O., Meyer, J. C., Erni, R., Rossell, M. D., Kisielowski, C., Yang, L., Park, C.-H., Crommie, M. F., Cohen, M. L., Louie, S. G., & Zettl, A. (2009). Graphene at the edge: stability and dynamics. *Science (New York, N.Y.)*, *323*, 1705–1708.

Hartelius, K., & Carstensen, J. (2003). Bayesian grid matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 162–173.

Ising, E. (1925). A contribution to the theory of ferromagnetism. *Z. Phys*, *31*, 253–258.

Kling, J., Vestergaard, J. S., Dahl, A. B., Stenger, N., Booth, T. J., Bøggild, P., Larsen, R., Wagner, J. B., & Hansen, T. W. (2014). Pattern recognition approach to quantify the atomic structure of graphene. *Carbon*, *74*, 363–366. doi:10.1016/j.carbon.2014.03.013.

Kotakoski, J., Krasheninnikov, A., Kaiser, U., & Meyer, J. (2011). From point defects in graphene to two-dimensional amorphous carbon. *Physical Review Letters*, *106*, 105505.

Kotakoski, J., Santos-Cottin, D., & Krasheninnikov, A. V. (2012). Stability of Graphene Edges under Electron Beam : Equilibrium Energetics versus Dynamic Effects. *ACS Nano*, *6*, 671–676.

Kubale, M. (2004). *Graph colorings* volume 349. American Mathematical Society.

Lindeberg, T. (1996). Scale-space: A framework for handling image structures at multiple scales. *CERN European Organization for Nuclear Research - Reports*, (pp. 27–38).

Meyer, J. C., Eder, F., Kurasch, S., Skakalova, V., Kotakoski, J., Park, H. J. J., Roth, S., Chuvilin, A., Eyhusen, S., Benner, G., Krasheninnikov, A. V., & Kaiser, U. (2012). An accurate measurement of electron beam induced displacement cross sections for single-layer graphene. *Physical Review Letters*, *108*, 196102.

Meyer, J. C., Geim, a. K., Katsnelson, M. I., Novoselov, K. S., Booth, T. J., & Roth, S. (2007). The structure of suspended graphene sheets. *Nature*, *446*, 60–3.

Meyer, J. C., Kisielowski, C., Erni, R., Rossell, M. D., Crommie, M. F., & Zettl, A. (2008). Direct imaging of lattice atoms and topological defects in graphene membranes. *Nano Letters*, *8*, 3582–3586.

Nolen, C., Teweldebrhan, D., & Denina, G. (2010). Large-Scale Automated Identification and Quality Control of Exfoliated and CVD Graphene via Image Processing Technique. *ECS Transactions*, *33*, 201–209.

Soille, P. (2003). *Morphological Image Analysis* volume 132. Springer.

Stadelmann, P. (2004). JEMS, EMS java version. *CIME-EPFL, Lausanne, Switzerland*, .

Wang, M., Liu, Q., Feng, J., Jiang, Q., Zou, X., & Pan, J. (2014). Recognition of defect structure of graphene by image processing technique. *Journal of Computational and Theoretical Nanoscience*, *11*, 391–395.

Wang, Z., Zhou, Y., Bang, J., Prange, M., Zhang, S., & Gao, F. (2012). Modification of Defect Structures in Graphene by Electron Irradiation: Ab Initio Molecular Dynamics Simulations. *The Journal of Physical Chemistry C*, *116*, 16070–16079.

Warner, J. H., Margine, E. R., Mukai, M., Robertson, A. W., Giustino, F., & Kirkland, A. I. (2012). Dislocation-driven deformations in graphene. *Science*, *337*, 209–212.

Welsh, D., & Powell, M. (1967). An upper bound for the chromatic number of a graph and its application to timetabling problems. *The Computer Journal*, *10*, 85–86.

Zhang, Y., Brar, V., Girit, C., Zettl, A., & Crommie, M. (2009). Origin of spatial charge inhomogeneity in graphene. *Nature Physics*, *5*, 722–726.

Acknowledgements

The Center for Nanostructured Graphene is sponsored by the Danish National Research Foundation, Project DNRF58. Financial support of the 7th Framework project GRAFOL is gratefully acknowledged. The A.P. Møller and Chastine Mc-Kinney Møller Foundation is acknowledged for their contribution toward the establishment of the Center for Electron Nanoscopy in the Technical University of Denmark. Thanks to Graphenea (San Sebastian, Spain) and Nicolas Stenger from DTU Fotonik for the provided graphene samples.

ESM 2. Temporal development

The two graphene samples have been imaged ten and twenty times respectively. The exposure time for each individual image is 1s. The structural changes in the graphene lattice over time is here analyzed by applying the proposed method to each exposure separately and extracting the estimated C-C bond lengths. Section 2.1 contains illustrations and descriptions for the ten exposures of the pristine graphene sample. Section 2.2 contains similar information for the twenty exposures of the altered graphene sample.

The hypothesis of whether the material changes significantly during the time of the exposures have been tested using a runs test. The null hypothesis of the average bond length changing randomly over time could not be rejected in neither the pristine case ($p = 0.81$), nor in the altered case ($p = 0.83$). Thus there is no evidence for a temporal trend in either case.

2.1. Pristine graphene

The ten exposures of the pristine graphene sample have undergone the analysis described in the main text. Figure 1 contains the cumulative distribution functions (CDFs) of the estimated C-C bond lengths for each exposure. The histograms representing the same distributions can be seen in the main text. The distributions are colored according to their number in the sequence, i.e., the first exposures are dark blue and the last exposures are bright green. It is apparent that the estimated distributions are very similar.

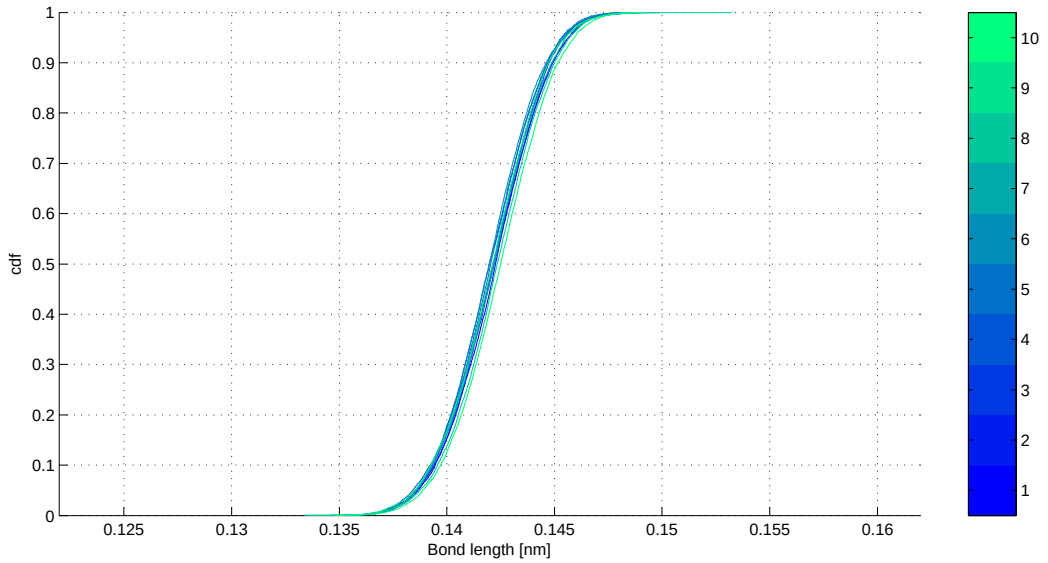


Figure 1: A cumulative distribution function (cdf) for each exposure. The color scale is such that the first exposure is dark blue and the last is bright green.

The first and the last exposure in the sequence have been overlaid the estimated C-C bonds in Figures 2–3. The C-C bonds are colored according to their length and clearly exhibit strong homogeneity over the extent of the image. Exposures two through nine exhibit similar structures.

A runs test is used to test the hypothesis of a temporal trend in the average C-C bond length. A runs test tests if the sequence of average bond lengths being above or below the overall average can be rejected to be random. As mentioned above, the hypothesis of random ordering of the average C-C bond lengths could not be rejected according to a runs test ($p = 0.71$). Thus, as expected from these visualizations, there is no evidence of a temporal trend.

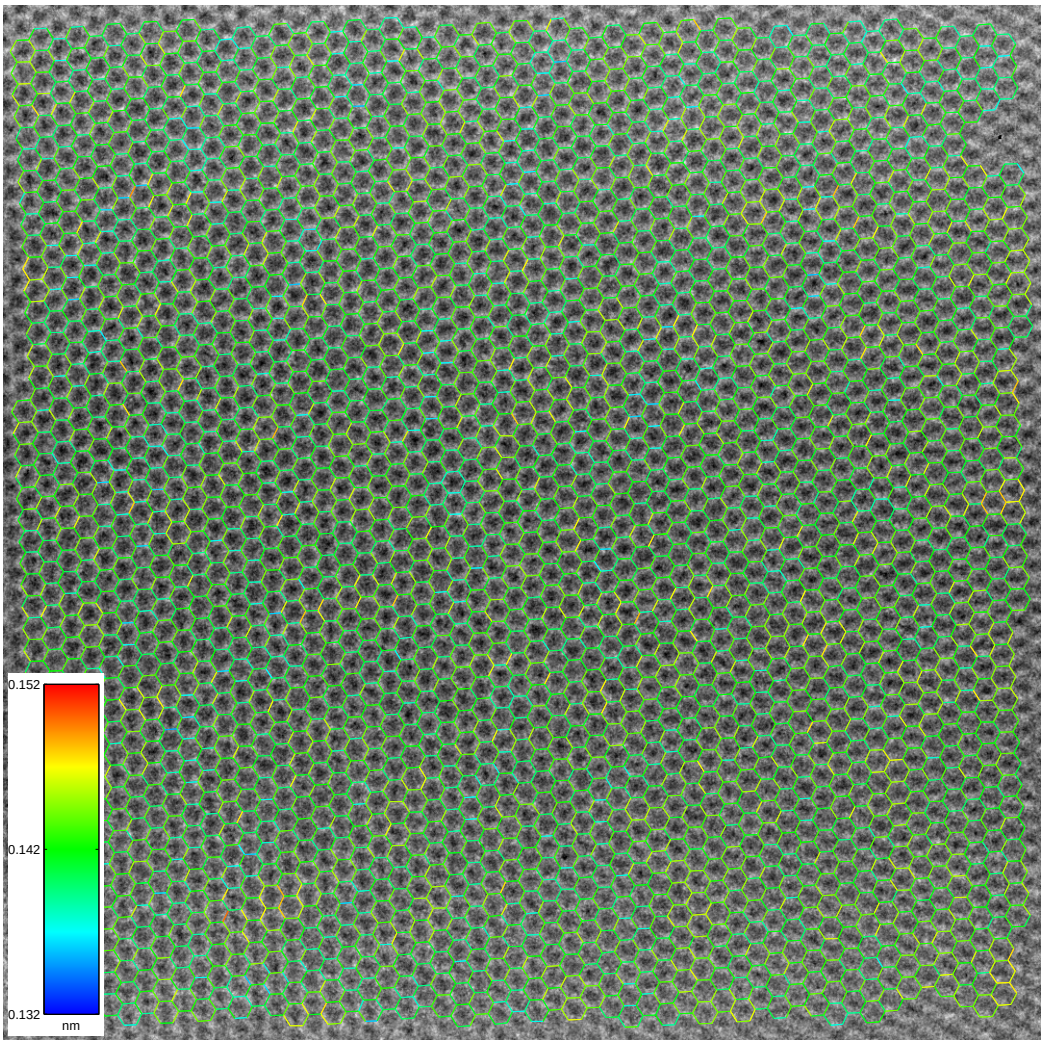


Figure 2: Exposure number 1. C-C bonds are colored according to length.

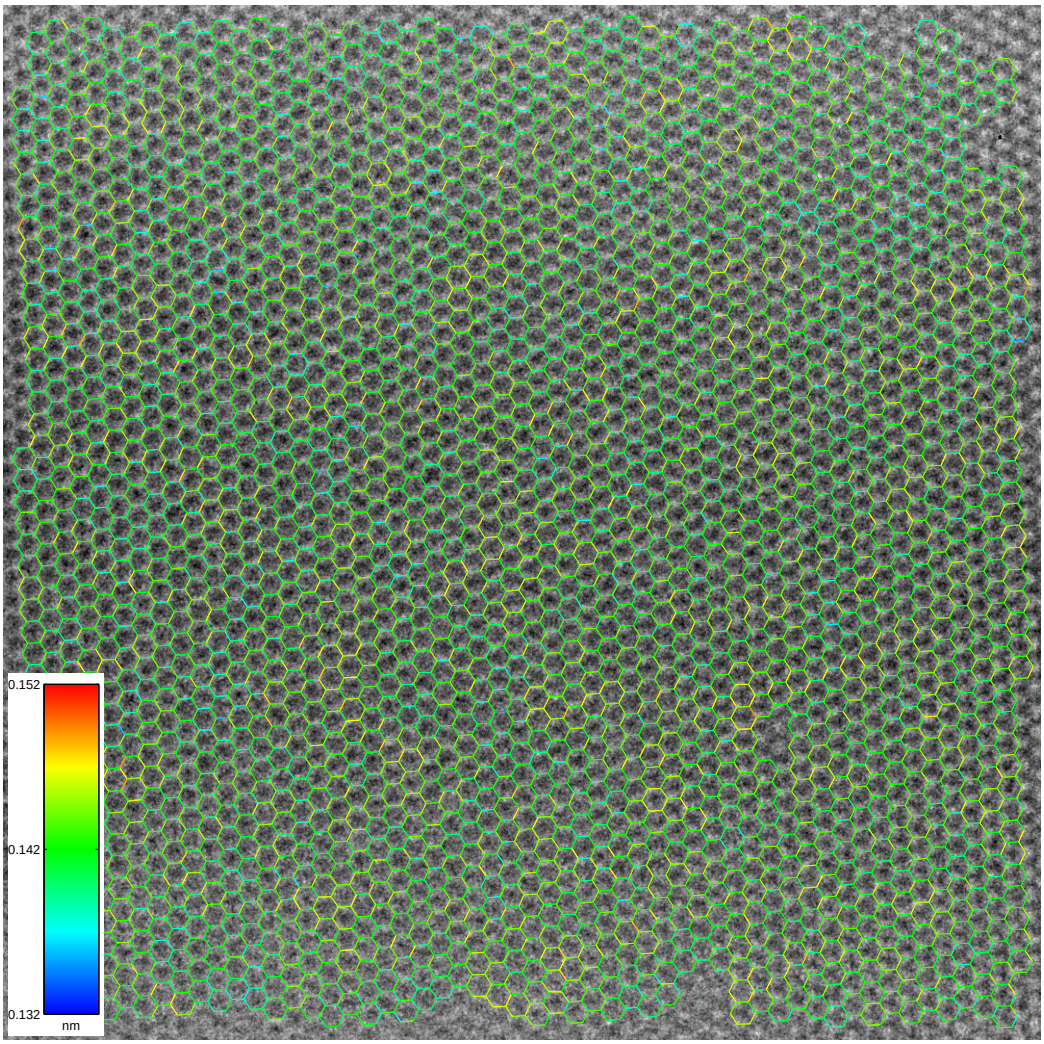


Figure 3: Exposure number 10. C-C bonds are colored according to length.

2.2. Altered graphene

The twenty exposures of the altered graphene sample have undergone the analysis described in the main text. Figure 4 contains the cumulative distribution functions (CDFs) for each of the twenty exposures. Inspecting this figure, there is certainly some variation between the exposures' distributions. As mentioned previously, the hypothesis of a random order of the average C-C bond lengths could not be rejected ($p = 0.96$), i.e., there is no evidence of a temporal trend.

The variation in the estimates can be inspected in Figures 5–8, where the first, eighth, fourteenth and twentieth exposures are overlaid the estimated structures as examples. It is seen that the areas to the left and the right of the hole are estimated relatively consistent, while the areas above and below the hole change in appearance from exposure to exposure. This local change is only evident in the overlay image and not in the distribution. Figures 6 and 7 show areas, where the method fails to confidently estimate any structure in very low contrast regions. These structural changes and the loss in contrast in some areas can be correlated to a beam induced effect as described in the main manuscript. The reduced contrast can originate from an offset in height, which results in a change in focus. This lack of estimated structure in some areas also adds to the observed variation.

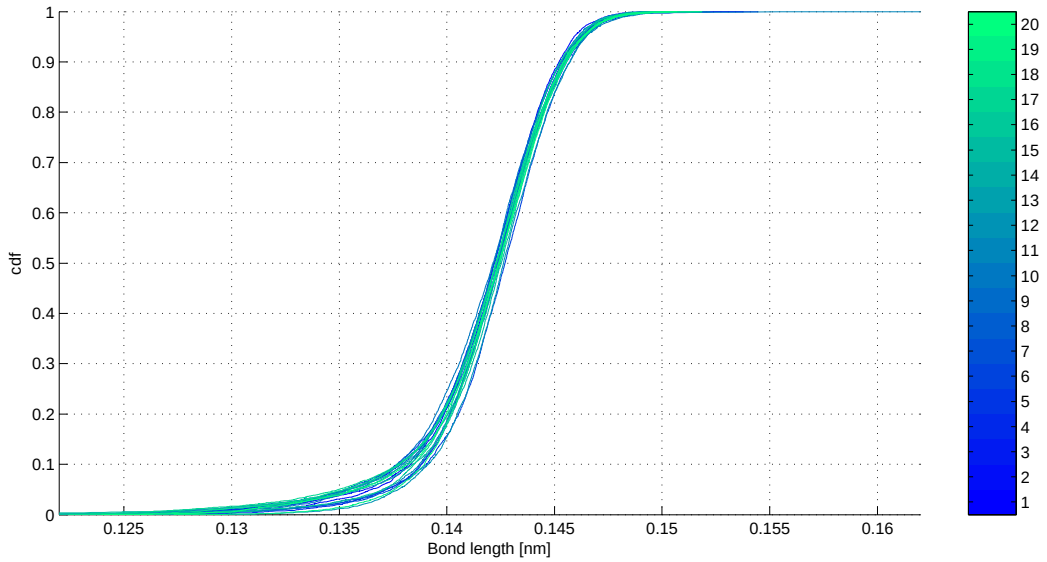


Figure 4: A cumulative distribution function (cdf) for each exposure. The color scale is such that the first exposure is dark blue and the last is bright green.

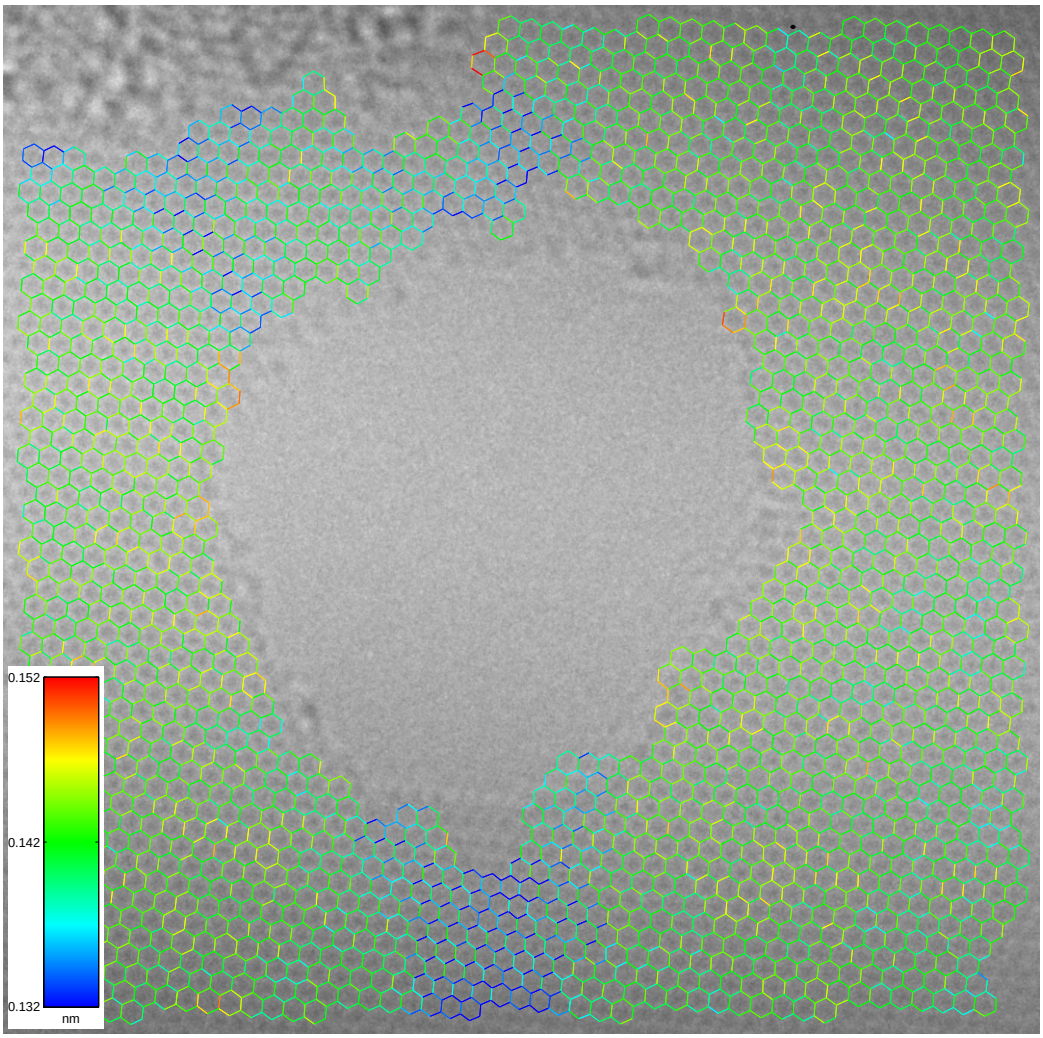


Figure 5: Exposure number 1. C-C bonds are colored according to length.

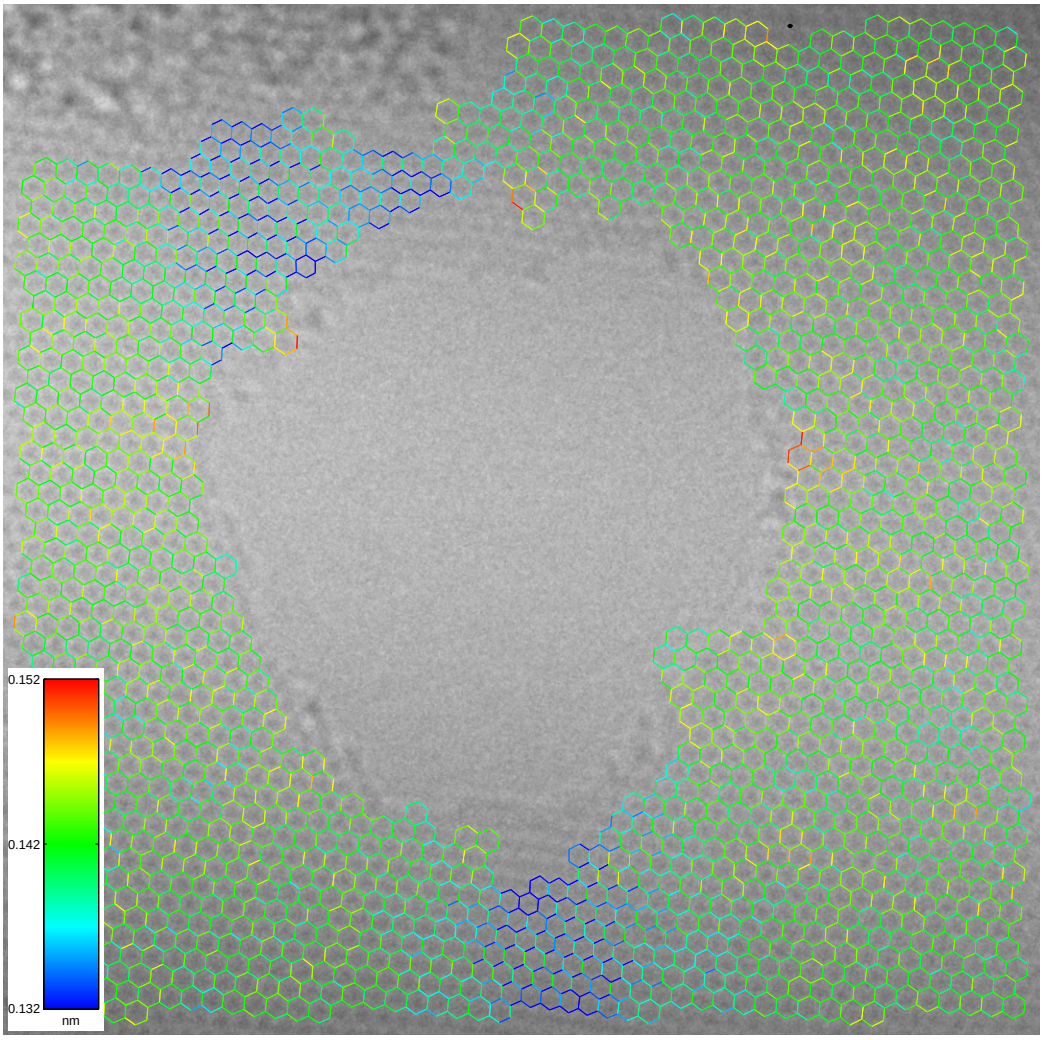


Figure 6: Exposure number 8. C-C bonds are colored according to length.

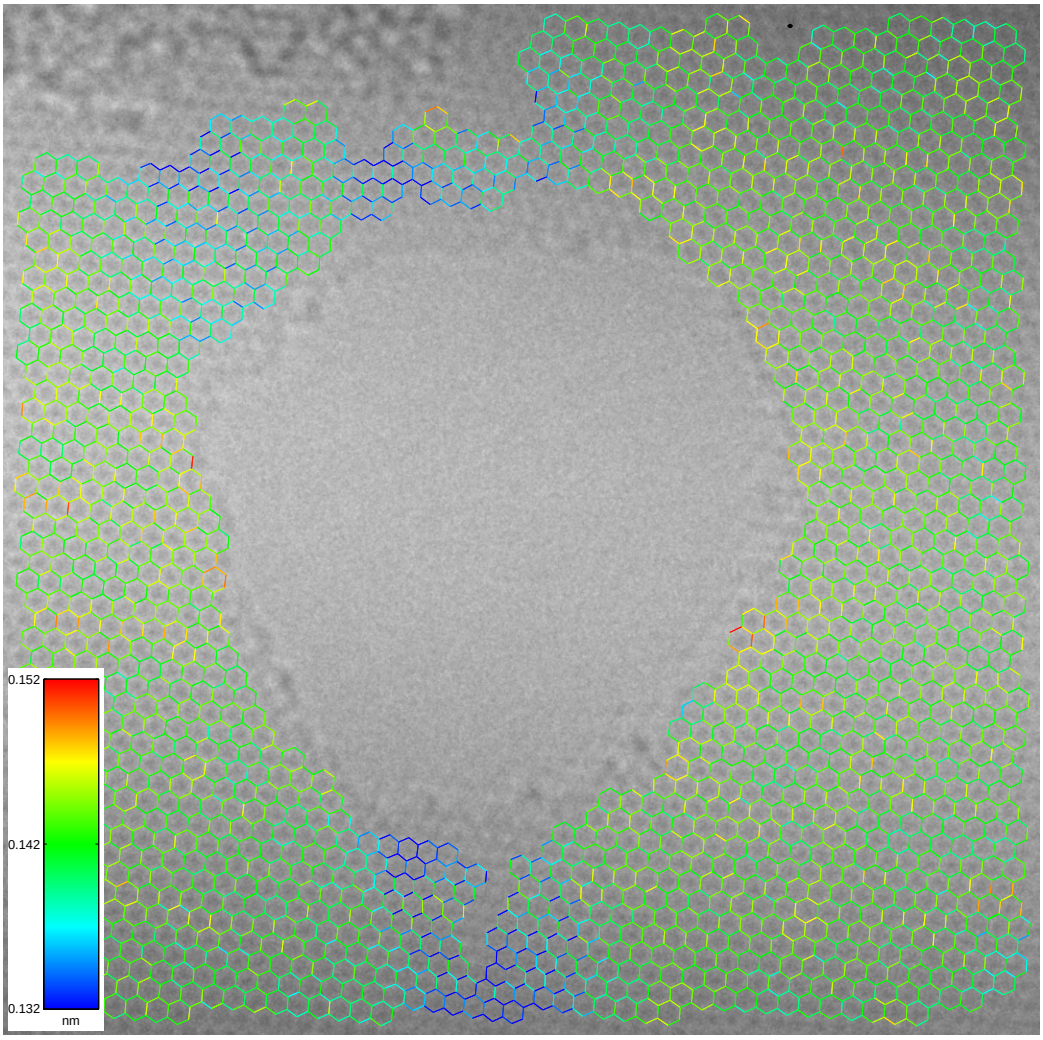


Figure 7: Exposure number 14. C-C bonds are colored according to length.

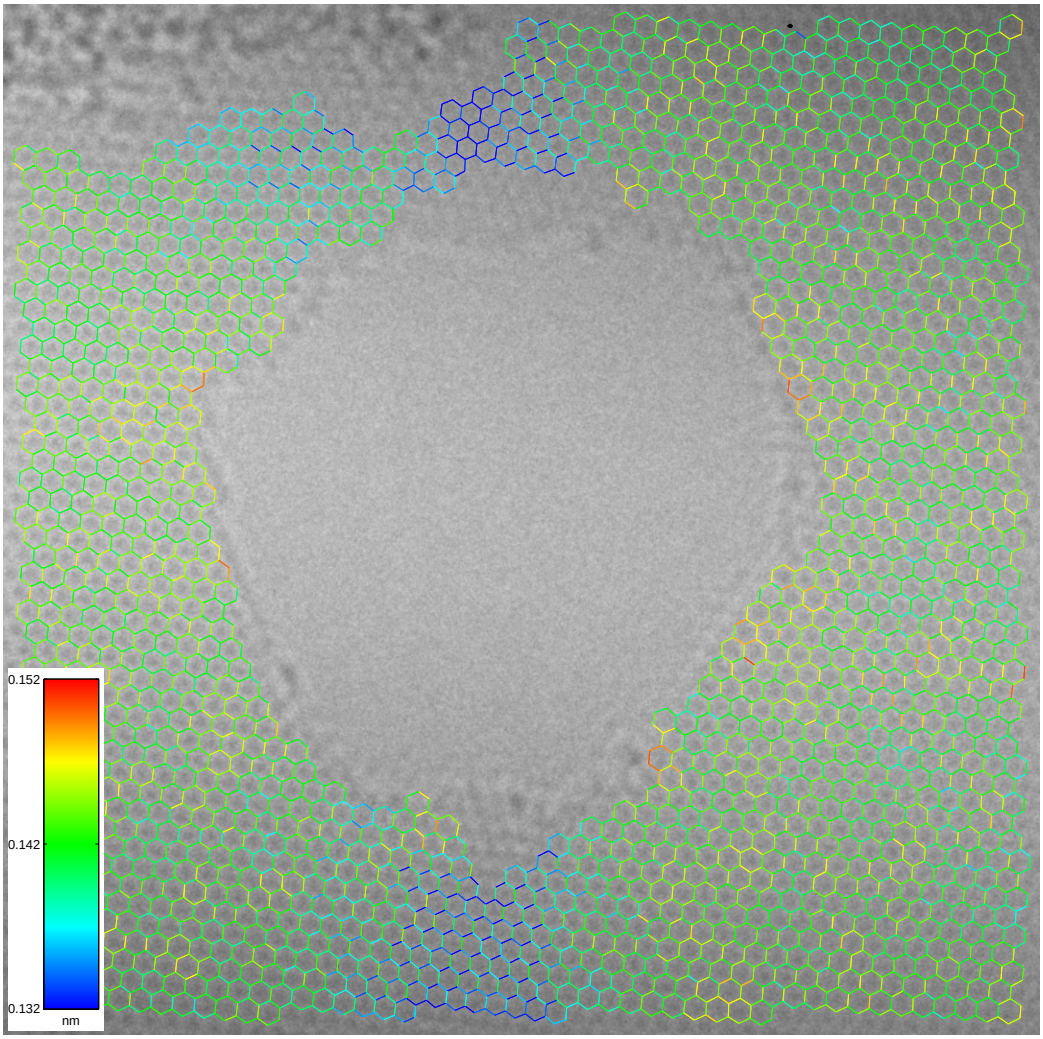


Figure 8: Exposure number 20. C-C bonds are colored according to length.

ESM 1. Simulation studies

Graphene structures with known C-C bond lengths are simulated to quantify the bias and precision of the presented algorithm under various degrees of strain and noise levels. First the simulation method is described, next the evaluation method and finally the bond length estimates are visualized and quantified.

HRTEM images are simulated using the software JEMS (Stadelmann, 2004). Multislice parameters were chosen according to the used FEI Titan with an energy spread of 0.3eV, negative Cs and positive defocus. As noise the preset "uniform noise" of the software was used with noise settings from 0 to 5%.

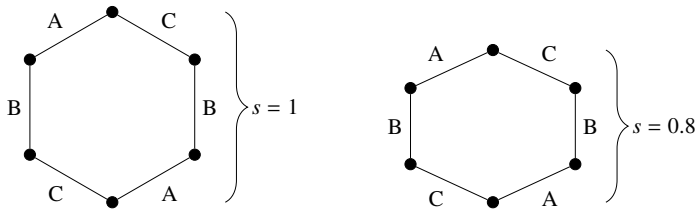


Figure 1: Illustration of the simulated strain by varying the scale parameter s. The lengths of the three pairs of parallel sides are denoted A, B and C. (left) A perfect lattice with no strain and (right) a lattice squeezed to $s = 0.8$ its original height.

The strain is simulated by scaling the vertical height of the unit cell with a factor s, illustrated in Figure 1. The lengths of each side will be referred to as A, B or C, i.e., the orientations according to this illustration. The lengths of these three sides, dependent on the scaling factor s, are:

$$A = C = \frac{1}{2} t \sqrt{3 + s^2} \tag{1}$$

$$B = st \tag{2}$$

where $t = 0.142$nm is the side length for a perfect hexagon with $s = 1$. The true values can be extracted from the simulation software and are listed in Table 1.

Simulation scenarios with four degrees of strain $s = \{1, 0.993, 0.986, 0.951\}$ and noise levels $\{0, \ldots, 5\}$ are analyzed, i.e., a total of 24 scenarios. Figures 2–4 show simulated structures with noise levels 1, 3 and 5 for the perfect lattice ($s = 1$). Figure 5 shows the most extreme straining with $s = 0.951$ and noise level 0. Note that the strain is not easily observed manually even though it is extreme for the material.

1.1. Evaluation and results

The grid structures are estimated from each of the simulated images using the method described in the main article. All parameters used for the grid structure estimation are equivalent to those in the main text.

The distributions of the estimated C-C bond lengths are illustrated using histograms and cumulative distribution functions (CDFs) in Figures 6–9. The CDF is useful when comparing distributions and does not depend on binning like the histogram. Each figure contains six plots: one histogram for each orientation and one CDF for each orientation. All six noise levels are shown as different colored lines in each plot. The vertical red line marks the expected bond length according to Eqs. (1) and (2). In all cases, it is noted that there is a high correspondence between the mode of the distribution and the expected bond length. For the noise free scenarions (level 0) only little mass of the distribution deviates from the expectation. For the most extreme case of strain ($s = 0.951$), the CDFs in Figures 9b, 9d and 9f show that a significant mass of each distribution is to the right of the expected bond length in the noise free scenario.

Figures 10– 13 show the estimates by overlaying the estimated C-C bonds on the original image colored according to length. For the three least degrees of strain, the estimate for noise level 3 is shown. The bond colors vary what seems randomly over the image for this noise level. However, the CDFs above tell us that on average, the bond lengths with similar orientations are also of similar length. This will also become evident from the numerical evaluation below. For the highest amount of strain, the no-noise case is shown (Figure 13). This color overlay clearly shows, that the strongest deviations from the expected bond lengths come from bonds at the image borders. There is an even more

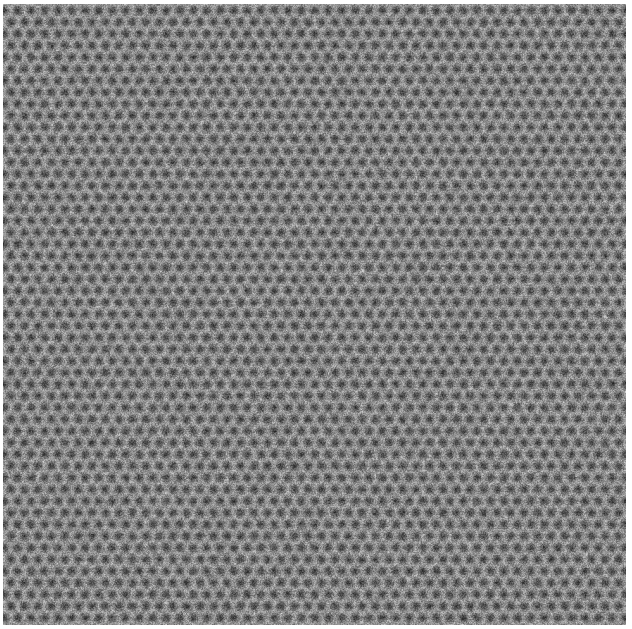


Figure 2: Perfect lattice ($s = 1$) with noise level 1.

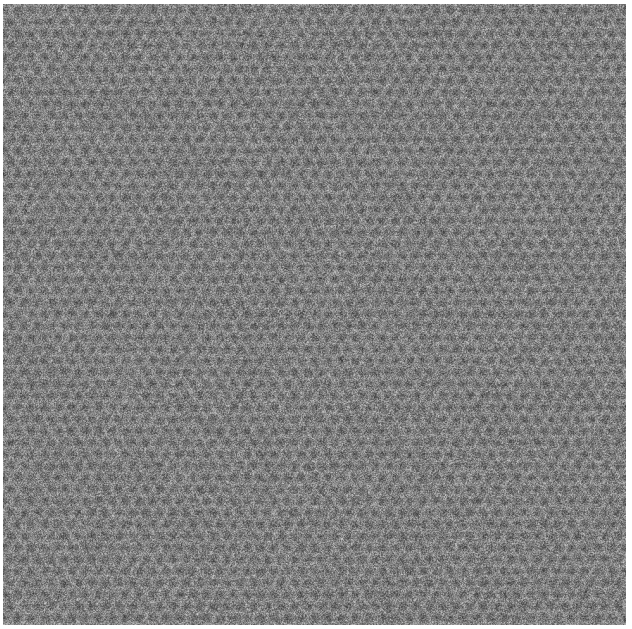


Figure 3: Perfect lattice ($s = 1$) with noise level 3.

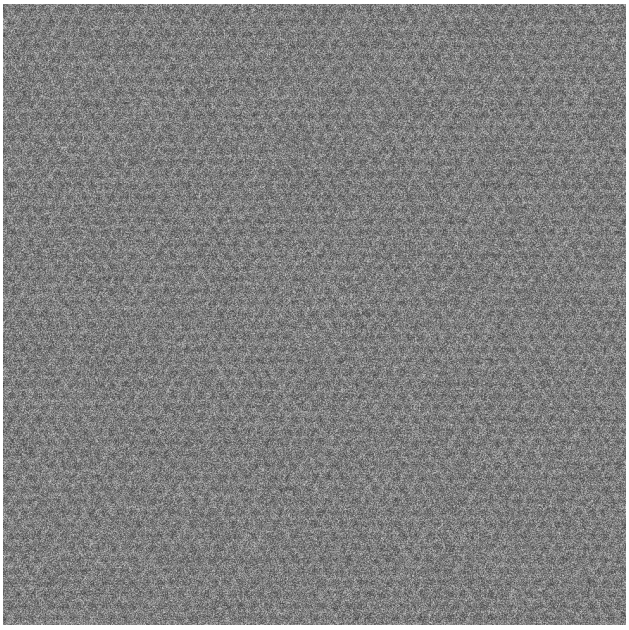


Figure 4: Perfect lattice ($s = 1$) with noise level 5.

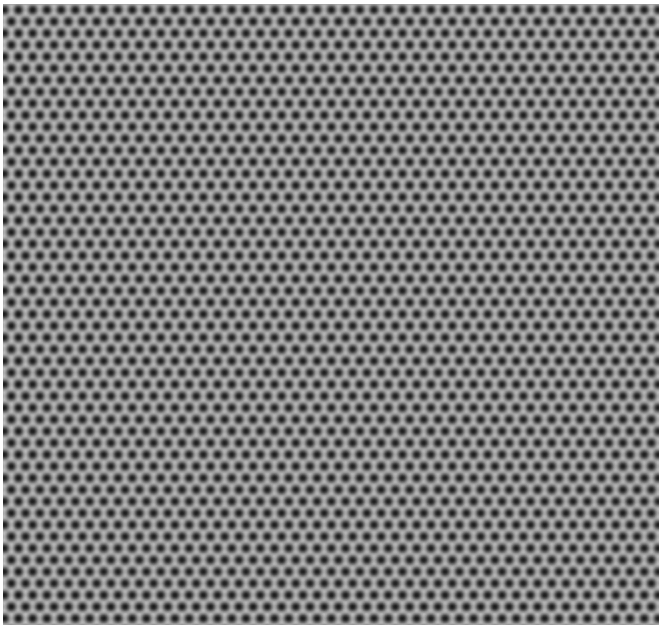


Figure 5: Lattice with simulated strain ($s = 0.951$) with no noise.

pronounced border effect, where bonds close the top and bottom image border are estimated to be longer than in the middle part, and vice versa for the left and right edges. This is what causes the "bump" in the CDF in Figure 9d.

In Table 1 the robustness of the method under these conditions is evaluated numerically in terms of bias and precision. A_{true}, B_{true} and C_{true} denote the known bond lengths under the given strain factor s according to Eqs. (1) and (2). The columns A_{err}, B_{err} and C_{err} contain the bias ± the precision in nm. The bias and precision for bonds of type A are calculated as

$$\text{bias} = \frac{1}{N_A}\sum_{i=1}^{N_A}\hat{A}_i - A_{\text{true}}$$

$$\text{precision} = \sqrt{\frac{1}{N_A}\sum_{i=1}^{N_A}\left(\hat{A}_i - A_{\text{true}}\right)^2},$$

where $\hat{A}_i$ is the estimate of the i'th C-C bond of N_A total bonds of type A. This is analogous for bonds of type B and C. ABC_{err} gives the bias and precision for all three orientations.

1.2. Conclusions

Based on inspection of the estimates of A, B and C in Table 1, the main conclusions to be drawn are:

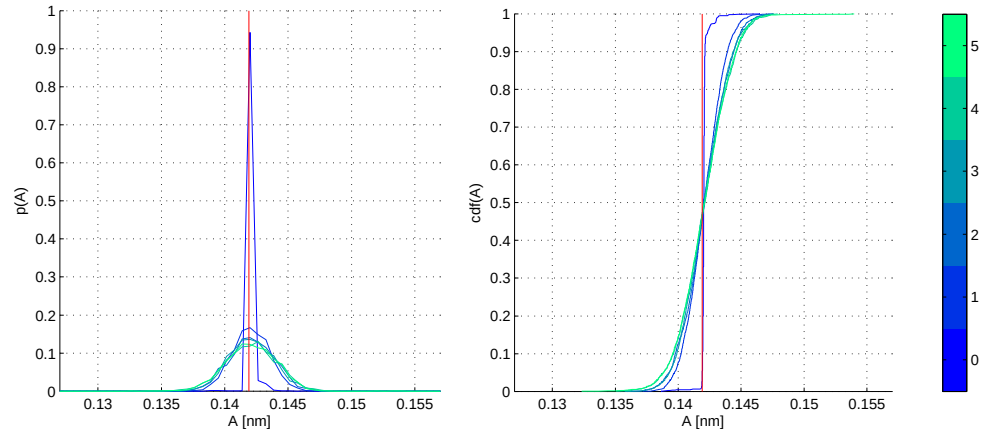
1. The lengths of C-C bonds under moderate strain ($s \in \{1, 0.993, 0.986\}$) and varying noise levels are estimated with low bias (≤ 0.0002nm) and a precision of ≤ 0.0020nm.
2. Under a simulation of heavy strain ($s = 0.951$) a maximum bias of 0.0004nm is obtained and a precision of ≤ 0.0025nm.
3. A bias of ≤ 0.0003nm and a precision of ≤ 0.0011 nm is achieved in simulations of noise free scenarios.
4. For a high amount of strain ($s = 0.951$) a border effect is apparent, yielding a non-homogeneous fit.

Considering the image resolution (≈ 0.0115nm/pixel), these results are very satisfactory. This sub-pixel precision is possible due to the large statistics within the image.

A possible alleviation of the border effect observed for $s = 0.951$ could be different choices of parameters for the grid structure estimation, e.g., a slower decrease of temperature in the simulated annealing scheme. Similarly, parameters could probably be tuned to improve the precision in many of the simulated scenarios. However, it is more meaningful to choose the same parameter settings for the simulation studies as for the real data in the main article.

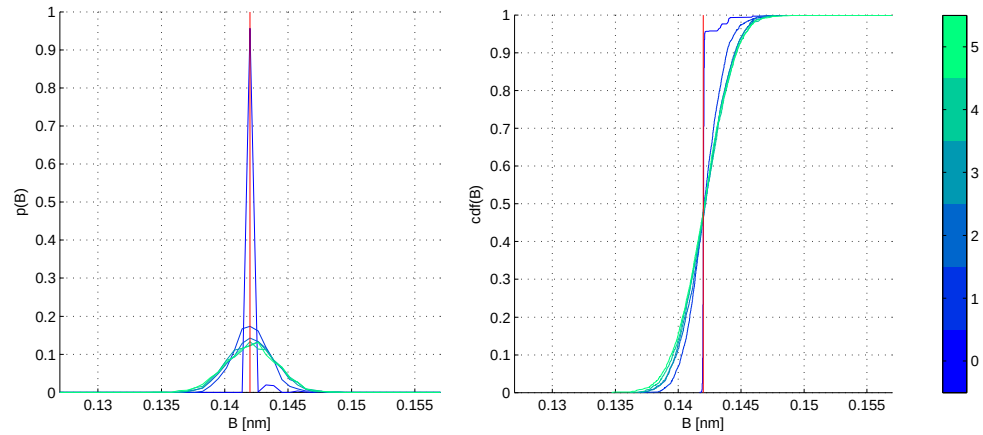
References

Stadelmann, P. (2004). JEMS, EMS java version. *CIME-EPFL, Lausanne, Switzerland*, .



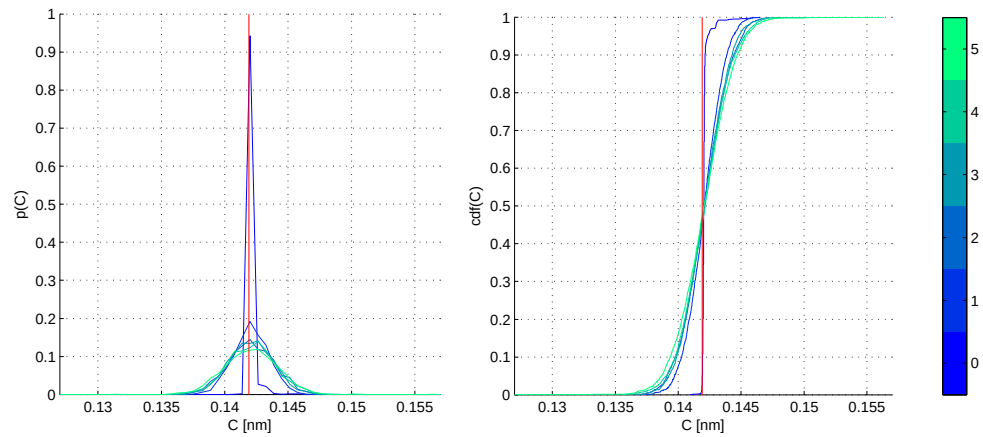
(a) Histogram for $\hat{A}$.

(b) CDF for $\hat{A}$.



(c) Histogram for $\hat{B}$.

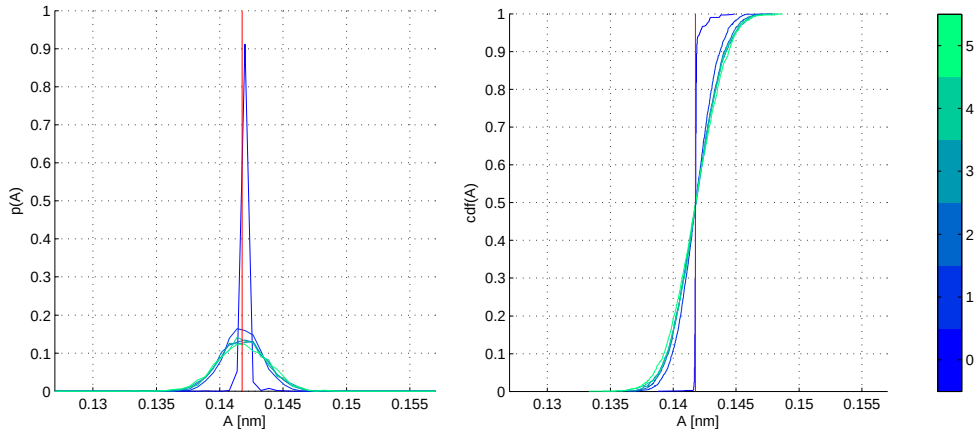
(d) CDF for $\hat{B}$.



(e) Histogram for $\hat{C}$.

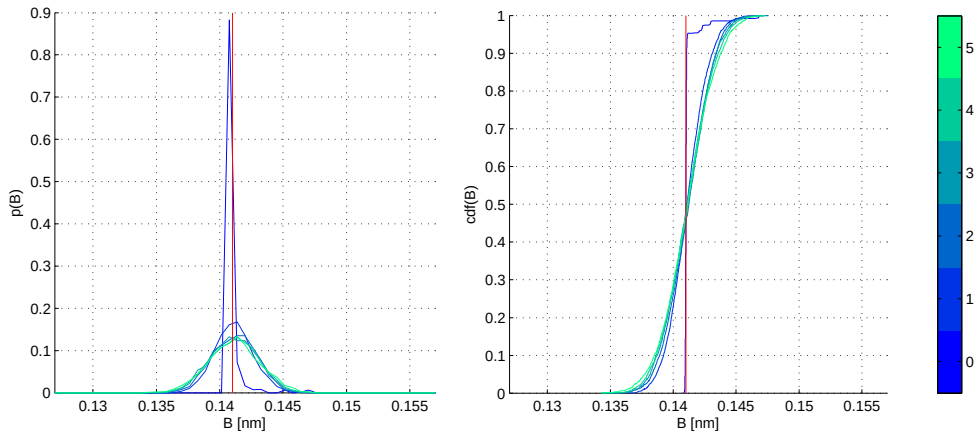
(f) CDF for $\hat{C}$.

Figure 6: Histograms and cumulative distribution functions (CDFs) for $\mathbf{s} = \mathbf{1}$. The distributions are colored according to the noise levels from 0–5. 65 equidistant bins in the range from 0.122 to 0.162 are used for the histograms. The vertical red line marks the known true bond length.



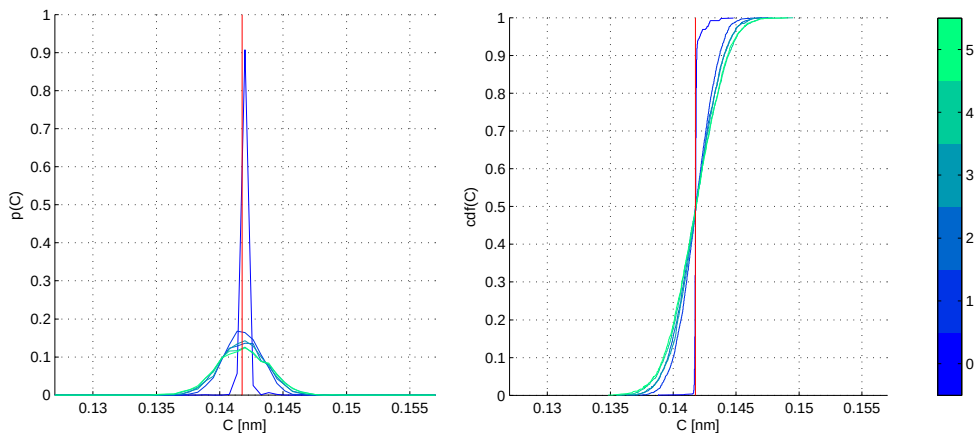
(a) Histogram for $\hat{A}$.

(b) CDF for $\hat{A}$.



(c) Histogram for $\hat{B}$.

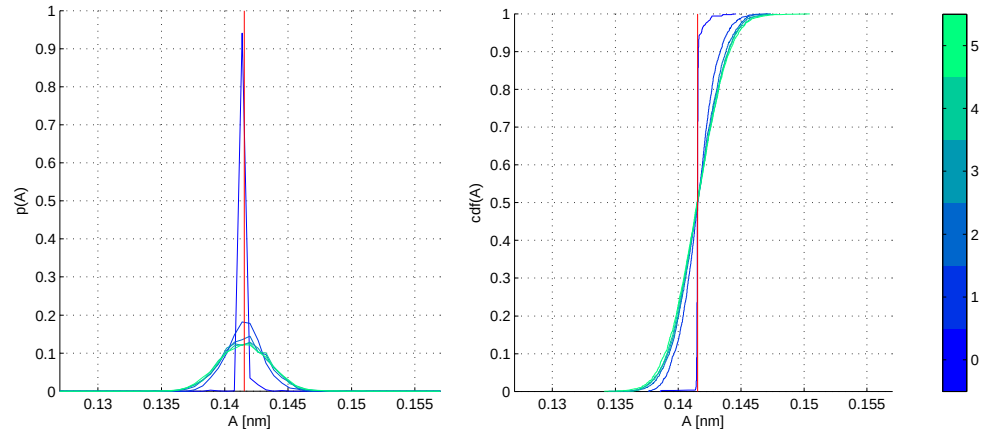
(d) CDF for $\hat{B}$.



(e) Histogram for $\hat{C}$.

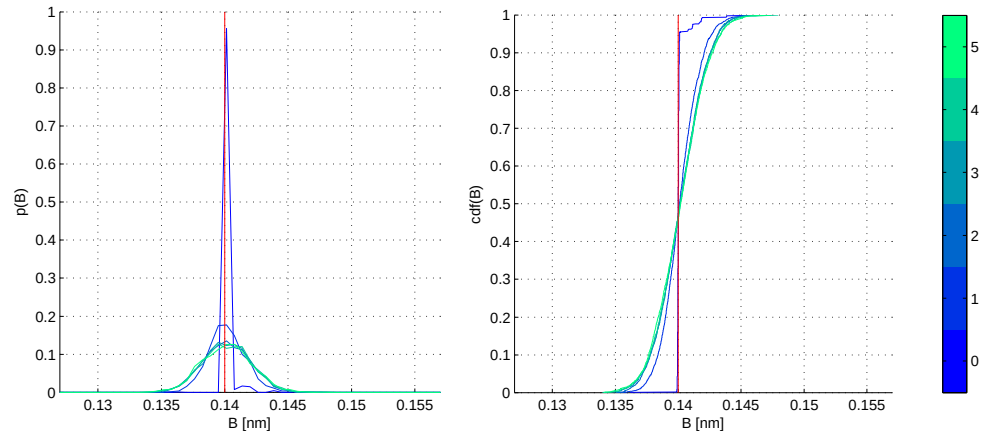
(f) CDF for $\hat{C}$.

Figure 7: Histograms and cumulative distribution functions (CDFs) for **s** = **0.993**. The distributions are colored according to the noise levels from 0–5. 65 equidistant bins in the range from 0.122 to 0.162 are used for the histograms. The vertical red line marks the known true bond length.



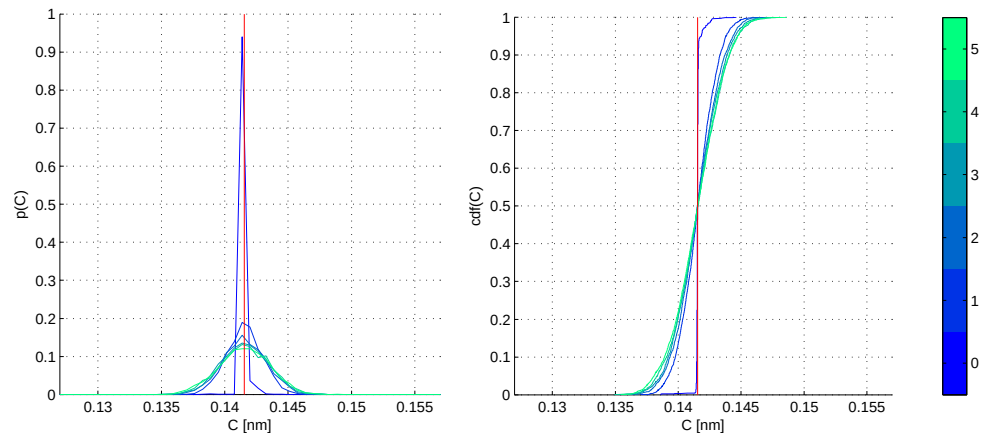
(a) Histogram for $\hat{A}$.

(b) CDF for $\hat{A}$.



(c) Histogram for $\hat{B}$.

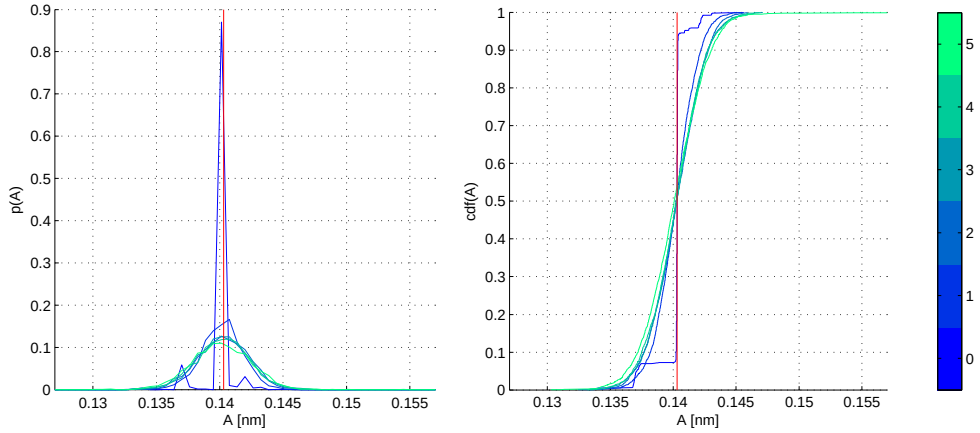
(d) CDF for $\hat{B}$.



(e) Histogram for $\hat{C}$.

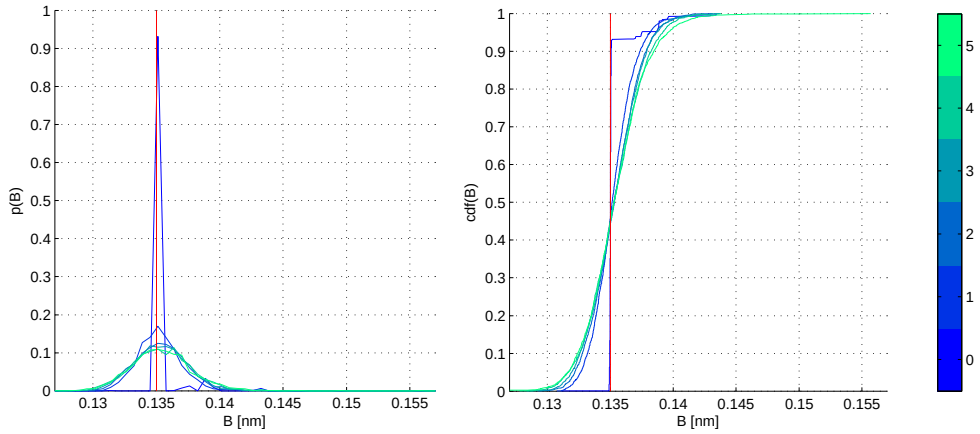
(f) CDF for $\hat{C}$.

Figure 8: Histograms and cumulative distribution functions (CDFs) for **s** = **0.986**. The distributions are colored according to the noise levels from 0–5. 65 equidistant bins in the range from 0.122 to 0.162 are used for the histograms. The vertical red line marks the known true bond length.



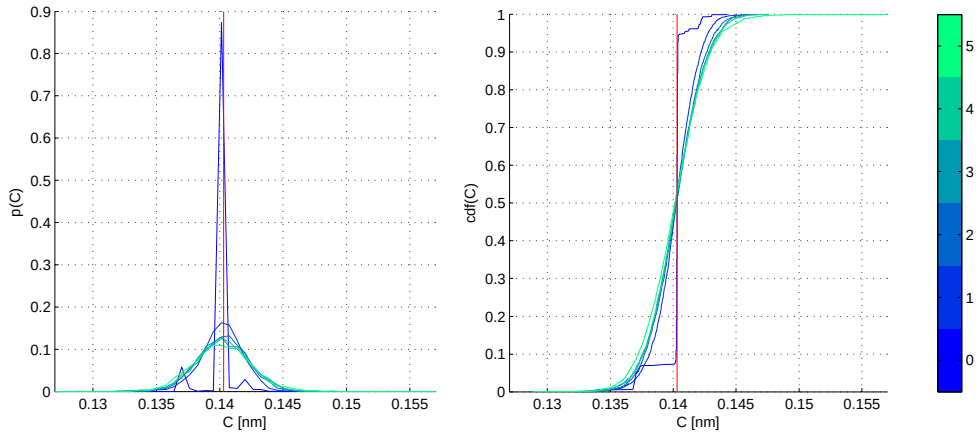
(a) Histogram for $\hat{A}$.

(b) CDF for $\hat{A}$.



(c) Histogram for $\hat{B}$.

(d) CDF for $\hat{B}$.



(e) Histogram for $\hat{C}$.

(f) CDF for $\hat{C}$.

Figure 9: Histograms and cumulative distribution functions (CDFs) for **s = 0.951**. The distributions are colored according to the noise levels from 0–5. 65 equidistant bins in the range from 0.122 to 0.162 are used for the histograms. The vertical red line marks the known true bond length.

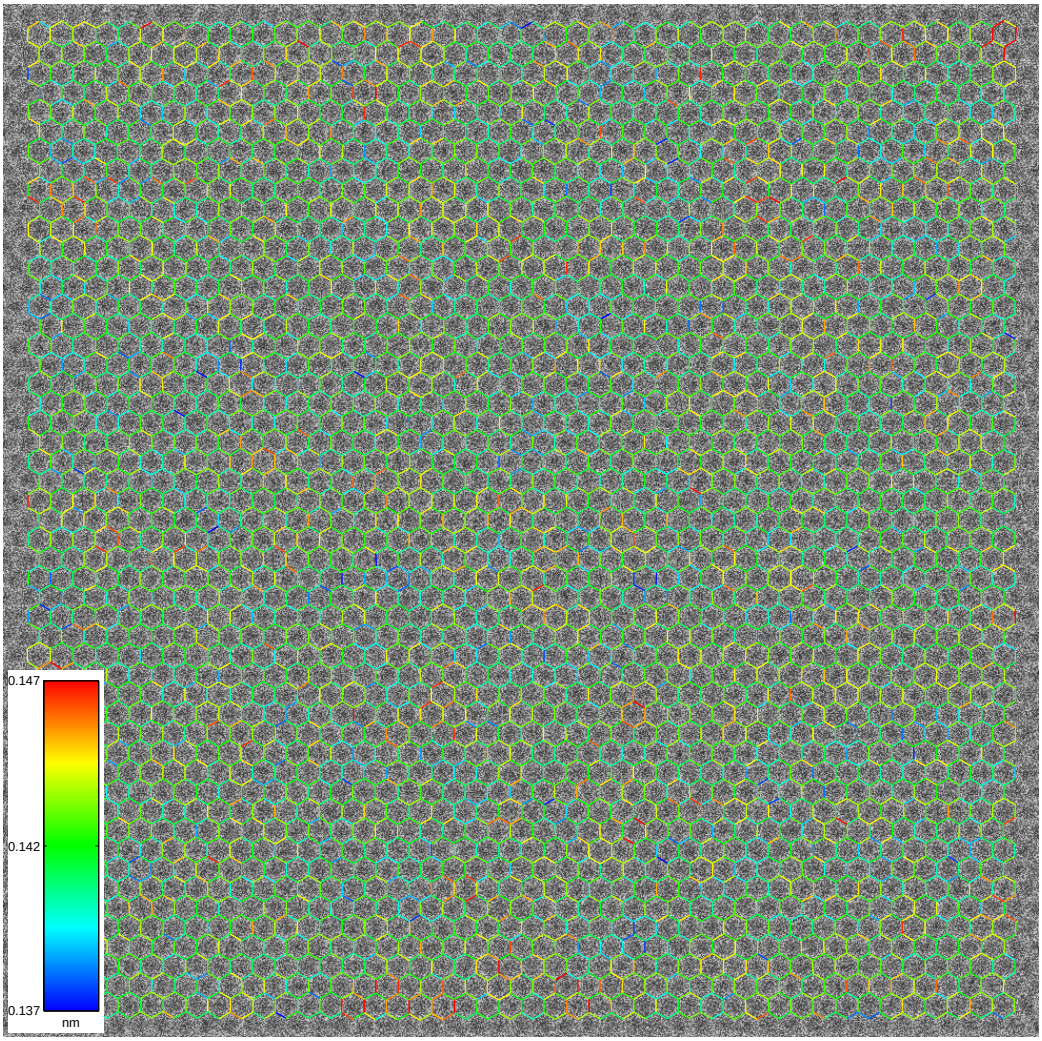


Figure 10: Estimated C-C bonds overlaid the simulated image with strain $\mathbf{s} = \mathbf{1}$ and noise level 3. C-C bonds are colored according to length.

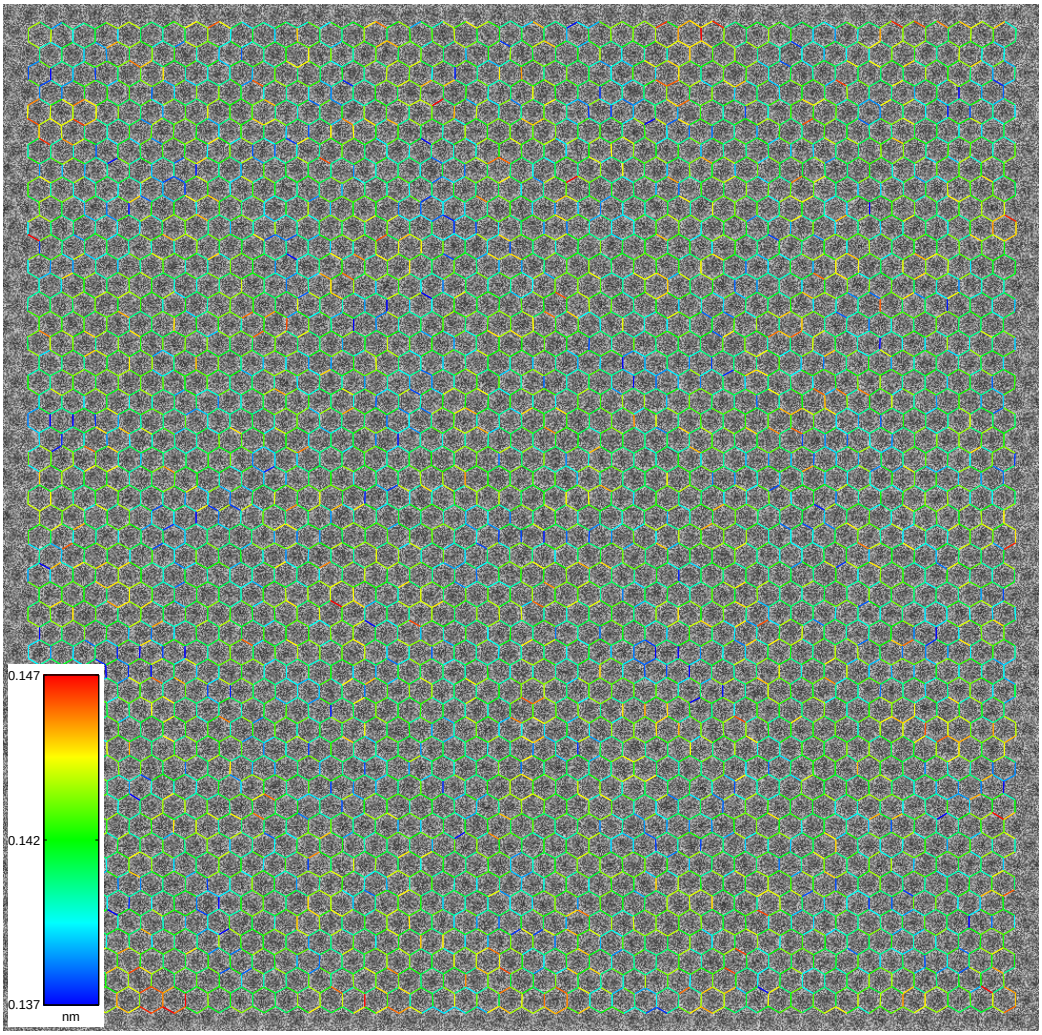


Figure 11: Estimated C-C bonds overlaid the simulated image with strain **s** = **0.993** and noise level 3. C-C bonds are colored according to length.

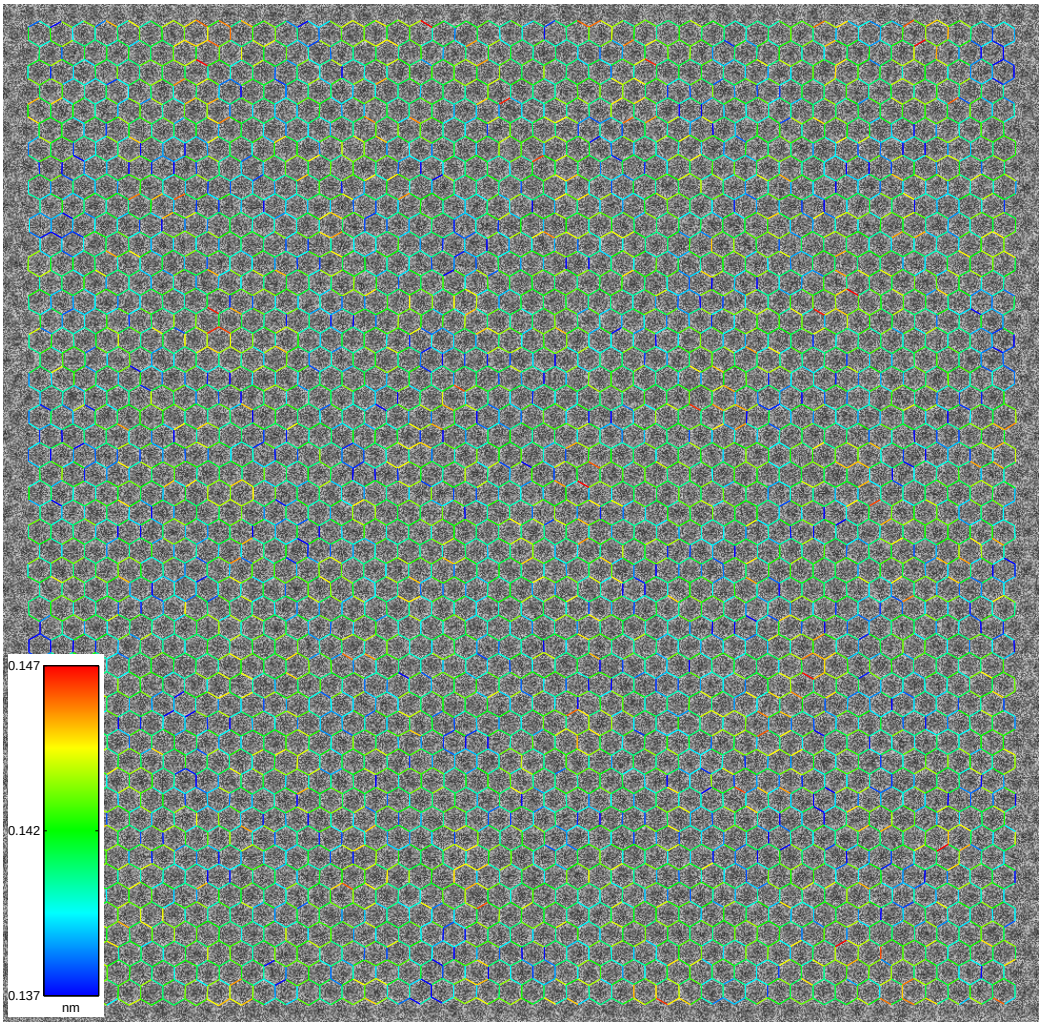


Figure 12: Estimated C-C bonds overlaid the simulated image with strain **s** = **0.986** and noise level 3. C-C bonds are colored according to length.

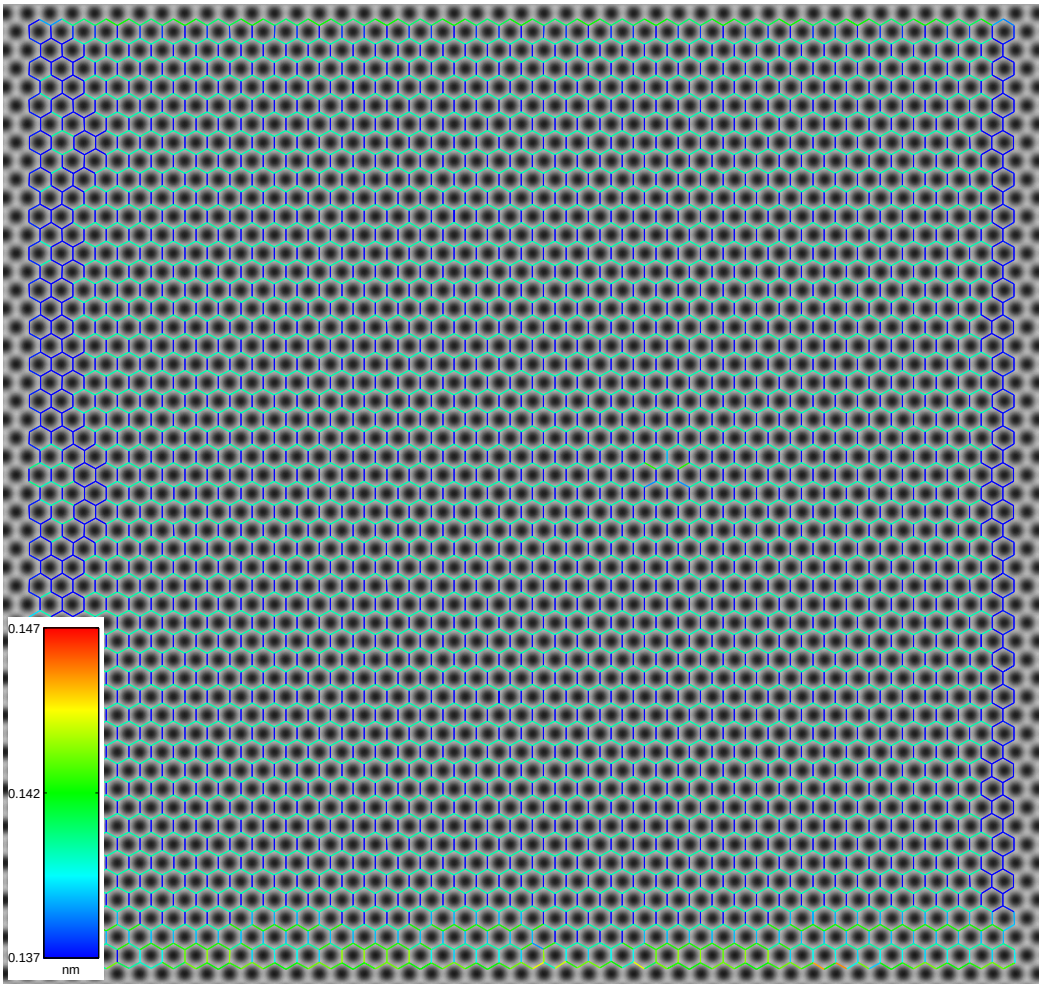


Figure 13: Estimated C-C bonds overlaid the simulated image with strain $\mathbf{s} = \mathbf{0.951}$ and no noise. C-C bonds are colored according to length.

Strain factor s	A_{true}	B_{true}	C_{true}	Noise	A_{err} [nm]	B_{err} [nm]	C_{err} [nm]	ABC_{err} [nm]
1.000	0.1419	0.1420	0.1419	0	0.0002 +/- 0.0003	0.0001 +/- 0.0005	0.0002 +/- 0.0003	0.0001 +/- 0.0004
				1	0.0002 +/- 0.0014	0.0001 +/- 0.0014	0.0002 +/- 0.0014	0.0001 +/- 0.0014
				2	0.0002 +/- 0.0017	0.0002 +/- 0.0018	0.0002 +/- 0.0018	0.0002 +/- 0.0018
				3	0.0002 +/- 0.0018	0.0001 +/- 0.0018	0.0002 +/- 0.0017	0.0002 +/- 0.0018
				4	0.0002 +/- 0.0020	0.0001 +/- 0.0019	0.0002 +/- 0.0019	0.0002 +/- 0.0019
				5	0.0002 +/- 0.0020	0.0001 +/- 0.0021	0.0002 +/- 0.0021	0.0001 +/- 0.0021
0.993	0.1418	0.1410	0.1418	0	0.0001 +/- 0.0003	0.0001 +/- 0.0006	0.0000 +/- 0.0003	0.0001 +/- 0.0004
				1	0.0000 +/- 0.0015	0.0001 +/- 0.0015	0.0000 +/- 0.0014	0.0001 +/- 0.0015
				2	0.0000 +/- 0.0017	0.0002 +/- 0.0017	0.0000 +/- 0.0017	0.0001 +/- 0.0017
				3	0.0000 +/- 0.0018	0.0001 +/- 0.0018	0.0001 +/- 0.0017	0.0001 +/- 0.0018
				4	0.0000 +/- 0.0018	0.0001 +/- 0.0019	0.0000 +/- 0.0020	0.0001 +/- 0.0019
				5	0.0000 +/- 0.0020	0.0001 +/- 0.0020	0.0000 +/- 0.0020	0.0000 +/- 0.0020
0.986	0.1415	0.1400	0.1415	0	0.0000 +/- 0.0003	0.0001 +/- 0.0004	0.0000 +/- 0.0003	0.0000 +/- 0.0003
				1	0.0000 +/- 0.0014	0.0002 +/- 0.0015	0.0000 +/- 0.0014	0.0001 +/- 0.0014
				2	0.0000 +/- 0.0018	0.0002 +/- 0.0018	0.0000 +/- 0.0017	0.0001 +/- 0.0018
				3	0.0000 +/- 0.0019	0.0002 +/- 0.0018	0.0000 +/- 0.0018	0.0001 +/- 0.0018
				4	0.0000 +/- 0.0019	0.0002 +/- 0.0019	0.0001 +/- 0.0019	0.0001 +/- 0.0019
				5	0.0000 +/- 0.0020	0.0002 +/- 0.0020	0.0000 +/- 0.0020	0.0000 +/- 0.0020
0.951	0.1403	0.1350	0.1403	0	-0.0002 +/- 0.0010	0.0003 +/- 0.0011	-0.0002 +/- 0.0010	0.0000 +/- 0.0011
				1	-0.0001 +/- 0.0016	0.0002 +/- 0.0017	-0.0001 +/- 0.0016	0.0000 +/- 0.0017
				2	-0.0001 +/- 0.0020	0.0003 +/- 0.0020	-0.0001 +/- 0.0019	0.0001 +/- 0.0020
				3	-0.0001 +/- 0.0020	0.0003 +/- 0.0021	-0.0001 +/- 0.0021	0.0000 +/- 0.0021
				4	-0.0001 +/- 0.0021	0.0004 +/- 0.0023	-0.0001 +/- 0.0021	0.0001 +/- 0.0021
				5	-0.0002 +/- 0.0024	0.0004 +/- 0.0025	-0.0002 +/- 0.0024	0.0000 +/- 0.0024

Table 1: Numerical evaluation of simulation studies. A_{true}, B_{true} and C_{true} denote the known bond lengths under the given strain factor s. The columns A_{err}, B_{err} and C_{err} contain the bias ± the precision in nm. The precision is the standard deviation of the estimates. ABC_{err} gives the bias and precision for all three orientations.

Paper F

Pattern recognition approach to quantify the atomic structure of graphene

CARBON 74 (2014) 363–366

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/carbon

Letter to the Editor

Pattern recognition approach to quantify the atomic structure of graphene

Jens Kling [a,e,*], Jacob S. Vestergaard [b], Anders B. Dahl [b], Nicolas Stenger [c,e], Tim J. Booth [d], Peter Bøggild [d,e], Rasmus Larsen [b], Jakob B. Wagner [a], Thomas W. Hansen [a,e]

[a] Center for Electron Nanoscopy (DTU Cen), Technical University of Denmark, Fysikvej 307, 2800 Kgs. Lyngby, Denmark
[b] Department of Applied Mathematics and Computer Science (DTU Compute), Technical University of Denmark, Matematiktorvet 303B, 2800 Kgs. Lyngby, Denmark
[c] Department of Photonics Engineering (DTU Fotonik), Technical University of Denmark, Ørsteds Plads 343, 2800 Kgs. Lyngby, Denmark
[d] Department of Micro- and Nanotechnology (DTU Nanotech), Technical University of Denmark, Ørsteds Plads 345E, 2800 Kgs. Lyngby, Denmark
[e] Center for Nanostructured Graphene (CNG), Technical University of Denmark, Ørsteds Plads 345E, 2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 19 December 2013
Accepted 8 March 2014
Available online 15 March 2014

ABSTRACT

We report a pattern recognition approach to detect the atomic structure in high-resolution transmission electron microscopy images of graphene. The approach provides quantitative information such as carbon–carbon bond lengths and bond length variations on a global and local scale alike.

Graphene is considered a key material for future electronic applications with the possibility of very high performance transistors [1], spintronics [2] and ballistic devices even at room temperature [3]. The degree to which the actual performance of graphene devices can live up to the theoretical predictions depends critically on the presence of defects or atomic configuration of edges. In essence, any deviations from perfect lattice periodicity can be important, which for instance is manifested in the sensitivity of electronic properties to strain [4].

Transmission electron microcopy (TEM) in general and high-resolution TEM (HRTEM) in particular can provide information about the atomic structure and defect landscape of graphene [5]. While important parameters like the carbon–carbon (C—C) bond length are possible to determine, this is usually done manually in small areas [6], due to time-consuming work of manually analyzing the HRTEM images. Here we describe a method for fast, automatic structure detection in graphene in a large number of sequentially acquired HRTEM images. The method enables quantitative information such as C—C bond length or bond length variations to be determined from images in a fast and reliable way, and can be used on many images to allow access to this information from a large area.

Suspended single-layer graphene synthesized by chemical vapor deposition (CVD) (Graphenea, Spain) or by mechanical

* *Corresponding author at:* Center for Electron Nanoscopy (DTU Cen), Technical University of Denmark, Fysikvej 307, 2800 Kgs. Lyngby, Denmark.
E-mail address: jenk@cen.dtu.dk (J. Kling).
http://dx.doi.org/10.1016/j.carbon.2014.03.013
0008-6223/

exfoliation of graphite [7] and transferred to TEM grids have been investigated using the automatic method. The graphene is imaged using a FEI Titan 80–300 Environmental TEM (ETEM) equipped with a monochromator at the electron gun and a spherical aberration (C_S)-corrector for the objective lens. All images are acquired with the microscope operated at 80 kV, which is below the knock-on threshold of carbon atoms in pristine graphene [8]. In order to optimize the imaging conditions and thereby the input for the structure detection, the electron beam energy spread was reduced to below 0.3 eV using the monochromator, while the C_S corrector was aligned to minimize the spherical aberration C_S. These conditions result in a resolution better than 0.12 nm, allowing us to resolve the 110-reflections of graphene and visualize the atomic structure accurately. The images are recorded using a Gatan US1000 CCD camera with an exposure time of 1 s.

The structure determination algorithm involves several steps. Utilizing Fourier transformation and local maxima detection, the mean graphene structure over the whole image is detected and used as starting point. The basis is a triangular lattice (triangulation) with a side length of roughly 0.247 nm, connecting three hexagonal centers (nodes) of the graphene structure. Hexagon center positions in the image are recognized as contrast extremes, minima for negative C_S or maxima for positive C_S. Nodes and triangles are removed from the triangulation when the local contrast properties or geometry deviate significantly from the expectations, and consequently areas like holes or amorphous material can be automatically omitted. In a final step, the node positions are adjusted by grid matching [9]. These steps enable a full reconstruction of the atomic structure of the observed graphene area in most cases. A more detailed description of the algorithm will be published elsewhere.

Figs. 1a and 2a show HRTEM images of two different areas of graphene. The hexagonal honeycomb lattice is easily recognized. Using negative C_S imaging, the carbon atoms are bright spots, with the centers of the hexagons appearing dark. In Fig. 1a, a hole in the graphene, formed under the influence of the electron beam is observed. Due to the inherent small signal-to-noise (S/N) ratio in the graphene image [5], as well as the continuous beam induced changes of the atomic structure at the edge [10], the termination of the hole cannot be completely resolved. Nevertheless, a predominant zigzag termination is assumed, which is in agreement with previous findings [11]. In Figs. 1b and 2b, the reconstructed graphene structure determined from the algorithm is overlaid, showing the actual hexagonal lattice of the graphene. The color coding represents the C—C bond lengths. As the absolute value retrieved from the images is dependent on the imaging conditions and the calibration, only the relative change within an image is considered. Fig. 1b, representing a pristine area, gives the impression of a homogenous distribution of bond lengths. This is reflected in the bond length histogram (Fig. 3 red) which exhibits a normal distribution.

For the case of a defective structure, as in Fig. 2, the algorithm detects nearly all graphene hexagons. The only exceptions are structures close to the hole and the edge termination itself, which most likely is due to insufficient imaging conditions, as mentioned above. The area of amorphous carbon from the transfer process or synthesis (top left)

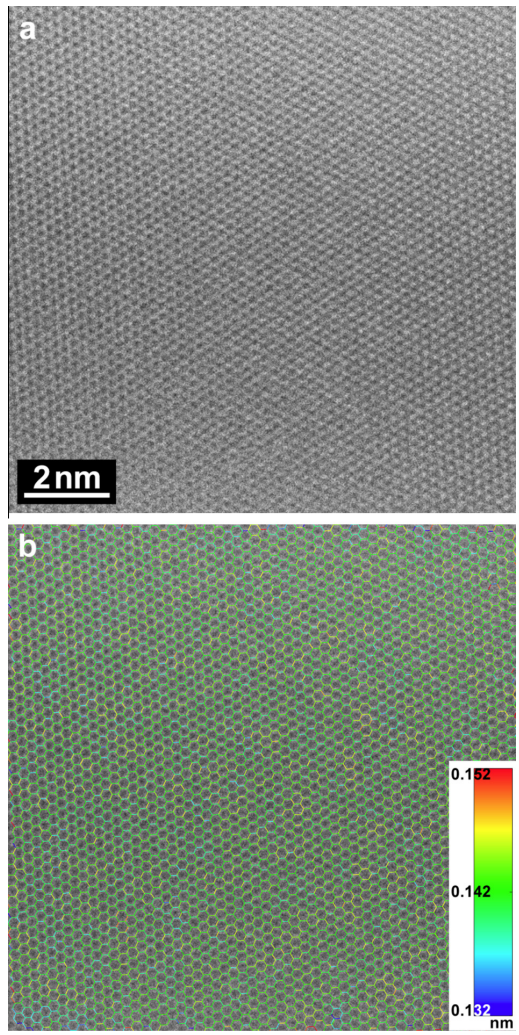


Fig. 1 – Pristine graphene. (a) HRTEM image, (b) image overlaid with the detected and reconstructed graphene structure. The color coding represents the C—C bond lengths. A homogenous distribution of bond lengths is observed.

was disregarded manually, but the area of the hole is detected by the algorithm and automatically left out in the analysis. A homogenous distribution of bond lengths is observed to the left and right of the hole. For the areas above and below the hole, significantly shorter bond lengths are detected. This is obvious in the histogram as well (Fig. 3 blue), where a tail towards shorter bond length is visible. Suspended graphene is known to form out-of-plane ripples [12], which are expected to have a lateral size of 2–20 nm and a height of 0.2–1 nm, with an inclination of ≈5° from the horizontal flat sheet. In this case, the projection of the bond length is measured from the image, making the lengths appear shorter. The shortest

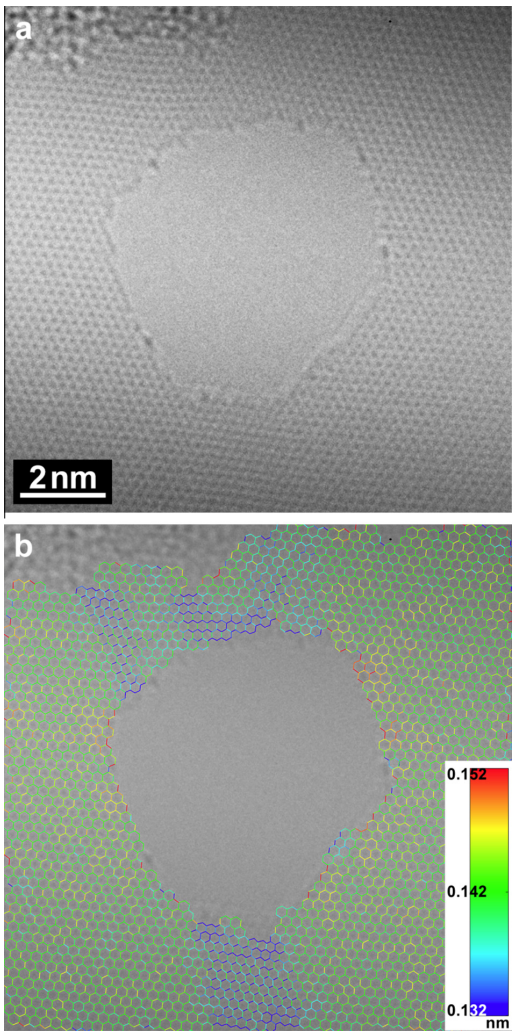


Fig. 2 – Graphene with a hole, formed under the influence of the electron beam. (a) HRTEM image, (b) image overlaid with the detected and reconstructed graphene structure. Color coding is the same as in Fig. 1b. Significantly shorter bond lengths with preferred orientation are observed above and below the hole.

observed bond lengths above and below the hole have a preferred orientation, almost horizontal in this image. This may be explained by the graphene sheet being slightly folded, as this should lead to a change in z-height as well as elastic deformation and strain mainly in the direction perpendicular to the fold, as we observe in the image. For an almost flat sheet, these bond lengths would represent a strain of about 7–8%; a pure inclination without bond length change would give an angle of about 22° between the two adjacent carbon atoms and an offset in height of about 0.05 nm. The rearrangement and knock-out of carbon atoms under the electron beam at the edge of the hole [10] can lead to structural deformation as well. A combination of both, a real shorting of the bond length and an artificial shortening due to projection in the image, is the most realistic explanation of our finding. It is important to notice, that even though the shorter bond lengths are represented in the global histogram, the local information, where these bonds actually appear in the structure, is only available in the image itself. The acquired local information in the bond length indicates a possible correlation between measured short bond lengths and folding of graphene. The nature of the bond length shortening will be clarified in a future study.

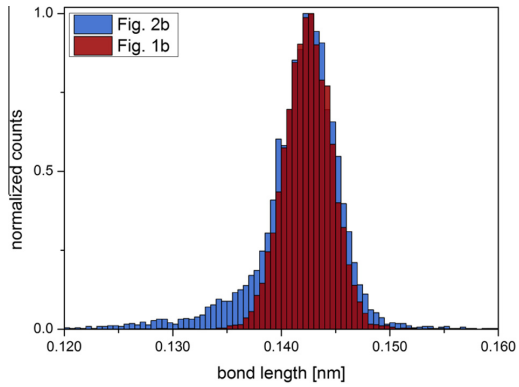


Fig. 3 – Histograms of the detected bond lengths from Fig. 1b (red) and Fig. 2b (blue). The bond lengths are binned by 0.5×10^{-3} nm steps; the counts are normalized for simpler shape comparison between the two histograms. The red histogram shows a normal distribution. The blue histogram exhibit a tail towards shorter bond length. (A color version of this figure can be viewed online.)

The automated structure detection for quantitative information extraction from high-resolution TEM images is possible for a large amount of images at relatively low time–cost and minimum manual interaction, making it easier and more feasible to follow structural changes in a series of images. Incremental and time dependent structural changes caused by either electron beam induced effects and/or external stimuli like temperature, current etc. can be monitored and quantified in future investigations in greater depths than standard image analysis procedures allow.

Acknowledgements

The Center for Nanostructured Graphene is sponsored by the Danish National Research Foundation, Project DNRF58. Financial support of the 7th Framework project “GRAFOL” is gratefully acknowledged. The A.P. Møller and Chastine Mc-Kinney Møller Foundation is acknowledged for their contribution toward the establishment of the Center for Electron Nanoscopy in the Technical University of Denmark. Thanks to Graphenea (San Sebastian, Spain) for providing the graphene sample.

N. Stenger acknowledges financial support by a Lundbeck Foundation Grant No. R95-A10663.

REFERENCES

[1] Schwierz F. Graphene transistors: status, prospects, and problems. Proc IEEE 2013;101:1567–84.
[2] Tombros N, Jozsa C, Popinciuc M, Jonkman HT, van Wees BJ. Electronic spin transport and spin precession in single graphene layers at room temperature. Nature 2007;448:571–4.
[3] Wang L, Meric I, Huang PY, Gao Q, Gao Y, Tran H, et al. One-dimensional electrical contact to a two-dimensional material. Science 2013;342:614–7.
[4] Pereira V, Castro Neto A. Strain engineering of graphene's electronic structure. Phys Rev Lett 2009;103:046801.
[5] Meyer JC, Kisielowski C, Erni R, Rossell MD, Crommie MF, Zettl A. Direct imaging of lattice atoms and topological defects in graphene membranes. Nano Lett 2008;8:3582–6.
[6] Warner JH, Margine ER, Mukai M, Robertson AW, Giustino F, Kirkland AI. Dislocation-driven deformations in graphene. Science (80-) 2012;337:209–12.
[7] Booth TJ, Blake P, Nair RR, Jiang D, Hill EW, Bangert U, et al. Macroscopic graphene membranes and their extraordinary stiffness. Nano Lett 2008;8:2442–6.
[8] Zobelli A, Gloter A, Ewels C, Seifert G, Colliex C. Electron knock-on cross section of carbon and boron nitride nanotubes. Phys Rev B 2007;75:245402.
[9] Hartelius K, Carstensen JM. Bayesian grid matching. IEEE Trans Pattern Anal Mach Intell 2003;25:162–73.
[10] Kotakoski J, Santos-Cottin D, Krasheninnikov AV. Stability of graphene edges under electron beam: equilibrium energetics versus dynamic effects. ACS Nano 2012;6:671–6.
[11] Girit CO, Meyer JC, Erni R, Rossell MD, Kisielowski C, Yang L, et al. Graphene at the edge: stability and dynamics. Science 2009;323:1705–8.
[12] Meyer JC, Geim AK, Katsnelson MI, Novoselov KS, Booth TJ, Roth S. The structure of suspended graphene sheets. Nature 2007;446:60–3.

Paper G

Classification of polarimetric SAR data using dictionary learning

Classification of polarimetric SAR data using dictionary learning

Jacob S. Vestergaard[a], Anders L. Dahl[a], Rasmus Larsen[a] and Allan A. Nielsen[b]

[a]Department of Informatics and Mathematical Modelling, Technical University of Denmark, Kgs. Lyngby, Denmark;
[b]National Space Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

ABSTRACT

This contribution deals with classification of multilook fully polarimetric synthetic aperture radar (SAR) data by learning a dictionary of crop types present in the Foulum test site. The Foulum test site contains a large number of agricultural fields, as well as lakes, wooded areas, natural vegetation, grasslands and urban areas, which makes it ideally suited for evaluation of classification algorithms.

Dictionary learning centers around building a collection of image patches typical for the classification problem at hand. This requires initial manual labeling of the classes present in the data and is thus a method for supervised classification. The methods aims to maintain a proficient number of typical patches and associated labels. Data is consecutively classified by a nearest neighbor search of the dictionary elements and labeled with probabilities of each class.

Each dictionary element consists of one or more features, such as spectral measurements, in a neighborhood around each pixel. For polarimetric SAR data these features are the elements of the complex covariance matrix for each pixel. We quantitatively compare the effect of using different representations of the covariance matrix as the dictionary element features. Furthermore, we compare the method of dictionary learning, in the context of classifying polarimetric SAR data, with standard classification methods based on single-pixel measurements.

Keywords: Discriminative dictionary learning, polarimetric SAR, multitemporal classification, Foulum

1. INTRODUCTION

Classification of crops using polarimetric SAR data is desirable due to the SAR's ability to operate under all weather conditions. The SAR measures dieletric and roughness properties of the target. A polarimetric SAR transmits and receives both horizontally and vertically polarized signals. From this different scattering properties of the target can be inferred. Classification of crops using these data relies on a difference in scattering properties between different types of crops.

Most previous work explicitly models the distribution of the scattering matrix or backscatter coefficients. This especially revolves around the complex Wishart distribution[1,2] and the Beta distribution.[3] No underlying distribution is assumed in the work presented here, similar to, e.g., entropy based approaches.[4]

While the application of multitemporal acquisitions has previously shown improved results over single-data acquisitions in classification crops,[5] due to the large interseasonal variations, the inclusion of a spatial context in the classification algorithms is not common in the literature. We quantitatively compare a spatially aware classification method with a standard maximum likelihood approach based on single-pixel measurements.

Dictionary learning for supervised image classification gathers a collection of typical patches for each class and ensures that these patches are separated in feature space. Thereby a sparse basis adapted to the problem at hand is built. This approach has previously shown success in texture classification,[6] biological[7] and geophysical applications.[8] The method will be further described in Section 3.

Further author information: (Send correspondence to J.S.V.)
J.S.V.: E-mail: jsve@imm.dtu.dk, Telephone: +45 45 25 33 51

Image and Signal Processing for Remote Sensing XVIII, edited by Lorenzo Bruzzone,
Proc. of SPIE Vol. 8537, 85370X · © 2012 SPIE · CCC code: 0277-786/12/$18
doi: 10.1117/12.974814

We apply this method to L-band single-polarimetric, dual polarimetry and full polarimetry SAR data from a subset of the Foulum data set[5,9] described in Section 2. The L-band data have previously been shown to be useful for classification of crops.[10] A quantitative comparison, varying the number of included temporal acquisitions and the size of the spatial neighborhood will be given in Section 4.

2. DATA

The data analyzed are multilook L-band fully polarimetric SAR data recorded over the Foulum test site. Polarimetric SAR data are acquired at four linear polarizations, HH, HV, VH, and VV, forming a scattering vector for the reciprocal case

$$\mathbf{k} = [S_{HH}\ S_{HV}\ S_{VV}]^T \tag{1}$$

where the subscripts denote receiving polarization before transmitting polarization. The data are multilooked for speckle reduction and represented in a covariance matrix

$$\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{k}_i\mathbf{k}_i^{*T} = \begin{bmatrix} \langle|S_{HH}|^2\rangle & \langle S_{HH}S_{HV}^*\rangle & \langle S_{HH}S_{VV}^*\rangle \\ \langle S_{HV}S_{HH}^*\rangle & \langle|S_{HV}|^2\rangle & \langle S_{HV}S_{VV}^*\rangle \\ \langle S_{VV}S_{HH}^*\rangle & \langle S_{VV}S_{HV}^*\rangle & \langle|S_{VV}|^2\rangle \end{bmatrix} \tag{2}$$

where spatial averaging is denoted by $\langle\rangle$. The elements of this covariance matrix can be described by nine independent real numbers,[3] namely the three real numbers on the diagonal and the real and imaginary parts of the three complex numbers above the diagonal. Thus a nine-vector $\mathbf{x}_{ij}$ represents the full polarimetric information for the (i,j)'th pixel in the acquired image

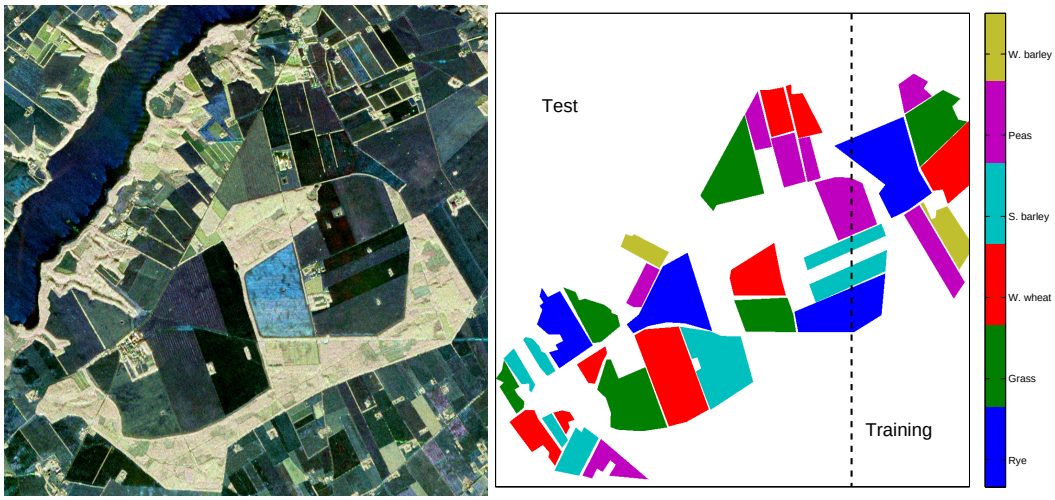
$$\mathbf{x}_{ij} = \begin{bmatrix} x_{ij}(1) \\ x_{ij}(2) \\ x_{ij}(3) \\ x_{ij}(4) \\ x_{ij}(5) \\ x_{ij}(6) \\ x_{ij}(7) \\ x_{ij}(8) \\ x_{ij}(9) \end{bmatrix} = \begin{bmatrix} \langle|S_{HH}|^2\rangle \\ \langle|S_{HV}|^2\rangle \\ \langle|S_{VV}|^2\rangle \\ \mathrm{Re}[\langle S_{HH}S_{HV}^*\rangle] \\ \mathrm{Im}[\langle S_{HH}S_{HV}^*\rangle] \\ \mathrm{Re}[\langle S_{HV}S_{VV}^*\rangle] \\ \mathrm{Im}[\langle S_{HV}S_{VV}^*\rangle] \\ \mathrm{Re}[\langle S_{HH}S_{VV}^*\rangle] \\ \mathrm{Im}[\langle S_{HH}S_{VV}^*\rangle] \end{bmatrix}. \tag{3}$$

Four different polarimetric modes are simulated by extracting different combinations of elements from this intensity vector. The number of elements in each mode is denoted p. Single-polarization in the horizontal direction consists of only the first element $\langle|S_{HH}|^2\rangle$ and similarly the vertical direction is represented by the third element, i.e., $p = 1$ for both. These modes are referred to as HH and VV respectively. Dual-copolarization (HHVV) is transmitting in both directions and receiving in both, though not acquiring the cross polarizations. Thus it consists of the $p = 4$ corner elements of the covariance matrix, corresponding to the first, third, eighth and ninth element of the intensity vector. The final mode is full polarimetry using all elements ($p = 9$) of the intensity vector for each pixel.

Figure 1a shows a pseudo-RGB image of the area analyzed, where red, green and blue are represented by the linearly stretched logarithmic values of the second, first and third element of the intensity vector in Eq. (3).

The Foulum test site contains 35 fields, where the grown crops are known, surrounded by a large area of unknown vegetation, lakes, grasslands and urban areas. The 35 fields are categorized into six classes of crops: rye, grass, winter wheat, spring barley, peas and winter barley. Figure 1b shows the ground reference data and the division into training and test set, which will be used in Section 4.

The full image is 1024×1024 pixels, where the first 700 columns are devoted to testing and the remainder for training. A few simple statistics on the training and test set can be seen in Table 1. It is worth noting that (1) the classes are not represented by an equal proportion of their occurrence in the test set, e.g., "grass" and "spring barley" are under represented in the training data, and (2) the total number of observations in each class is far from equal. Four temporally separate acquisitions covering this area are considered in Section 4.



(a) (b)

Figure 1: a) Pseudo-RGB image of the fully polarimetric SAR data. Red, green and blue represents the second, first and third element of the intensity vector in Eq. (3). b) Ground reference data for the Foulum test site consist of six known classes of crops and a large area of unknown background.

Class	# of fields	N_m	N_m/N^{total}	N_m^{train}/N_m	N_m^{test}/N_m
Rye	4	61583	0.23	0.43	0.57
Grass	6	56836	0.21	0.2	0.8
W. wheat	8	57685	0.22	0.18	0.82
S. barley	7	39481	0.15	0.14	0.86
Peas	8	42846	0.16	0.36	0.64
W. barley	2	9622	0.04	0.62	0.38

Table 1: Simple statistics on the division of classes into training and test sets. N_m is the number of observations in the m'th class. The division is illustrated in Figure 1b.

3. METHODS

We propose using a learned discriminative dictionary of polarimetric SAR data patches and extend these to include multiple temporal acquisitions. This method will be described below. A Bayesian maximum likelihood classifier is used for comparison and will be described in Section 3.2.

3.1 Discriminative dictionary learning

Often texture contains important information for image segmentation. This is utilised in the segmentation approach based on discriminative image patches.[6] The segmentation is done in small image patches of $\sqrt{n} \times \sqrt{n}$ pixels using a learned intensity dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$. Each image patch is concatenated to form a vector of size n and the dictionary contains m of these vectors. In addition to the intensity dictionary a label dictionary $\mathbf{L} \in \mathbb{R}^{nl \times m}$ is given where the i'th column vector in $\mathbf{L}$ correspond to the i'th column vector in the intensity dictionary $\mathbf{D}$. The label dictionary vectors are concatenated from image patches of $\sqrt{n} \times \sqrt{n} \times l$ pixels with a channel for each l class labels. The pixel values of the label patches correspond to the probability of a given label.

Segmentation is performed using a nearest neighbour classification among the column vectors of the intensity dictionary. For the image to be segmented a patch $\mathbf{x}$ of $\sqrt{n} \times \sqrt{n}$ pixels is chosen. The nearest neighbour $\mathbf{d}_j$ is found by

$$\mathbf{d}_j = \arg\min_{\mathbf{D}} ||\mathbf{d}_i - \mathbf{x}||, \quad i \in \{1, ..., m\}. \tag{4}$$

The corresponding label vector $\mathbf{l}_j$ contains the probabilities for the label classes, and choosing the most probable label for each pixel will provide a segmentation for that image patch. Segmenting the entire image is done by densely sampling overlapping image patches of $\sqrt{n} \times \sqrt{n}$ and averaging the overlapping regions. Hereby an image containing label probabilities is obtained.

The segmentation procedure is supervised and need training samples to build the dictionaries. Given a set of training image patches and corresponding label patches the dictionary is constructed using a weighted k-means clustering approach. The weights are obtained from the label patches in such a way that image patches in a cluster has similar label patches. Details on building the dictionary can be found in Dahl and Larsen.[6]

3.1.1 Multitemporal dictionary atoms

For the purposes here we extend the dictionary atoms to include multitemporal acquisitions of polarimetric SAR scatter information. This is done by concatenating all information to a vector for each pixel, as illustrated in Figure 2. For a dictionary atom of spatial extent $\sqrt{n} \times \sqrt{n}$ pixels, p unique elements from the intensity vector in Eq. (3) and Δt temporal acquisitions, the number of features – and thereby the dimensionality of the intensity dictionary – is $n \cdot p \cdot \Delta t$.

The dimensions of the dictionary thus rapidly grows when including a larger spatial context, working in a more complex polarimetric mode or including more temporal acquisitions.

3.2 Maximum likelihood classification

The classification results obtained by employing discriminative dictionary learning are compared with the standard Bayesian Maximum Likelihood (ML) classifier. For multilook single-polarimetric SAR data it is assumed that the backscatter coefficients follow a Gaussian distributions, when the number of looks is large enough. We make this assumption here.

The negative log-likelihood $L_m(\mathbf{f})$ for the feature vector $\mathbf{f}$ belonging to the m'th class with covariance $\mathbf{\Sigma}_m$ and mean $\boldsymbol{\mu}_m$ is

$$L_m(\mathbf{f}) = \frac{1}{2} (\mathbf{f} - \boldsymbol{\mu}_m)^T \mathbf{\Sigma}_m^{-1} (\mathbf{f} - \boldsymbol{\mu}_m) + \frac{1}{2} \log |\mathbf{\Sigma}_m| - \log \pi_m \tag{5}$$

according to the ML classifier. Here π_m denotes the prior probability of the m'th class. We assume equal prior probabilities for all classes. The feature vector $\mathbf{f}$ is of length $\Delta t \cdot p$ with the same definitions as previously.

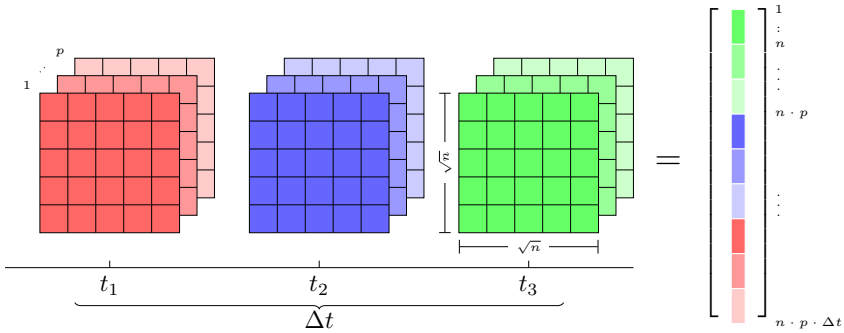


Figure 2: Illustrating concatenation of multitemporal (here three) acquisitions of spatially extending polarimetric scattering information into vector form. $\sqrt{n}$ is the width of the dictionary atom, p the number of included elements from the vector in Eq. (3) and Δt is the number of temporal acquisitions included.

Similarly, the dual-copolarimetric and fully polarimetric SAR data are assumed to follow a complex Wishart distribution.[1,5] For these modes the ML classifier takes the form

$$L_m(\mathbf{x}) = n\left[\mathrm{Tr}(\bar{\mathbf{\Sigma}}_m^{-1}\mathbf{x}) + \log|\bar{\mathbf{\Sigma}}_m|\right] - \log\pi_m \tag{6}$$

where $\mathbf{x}$ is the covariance matrix for the observation to be classified, n is the number of looks, and $\bar{\mathbf{\Sigma}}_m$ is the average covariance matrix for all observations belonging to class m. For dual-copolarization data this is a complex matrix of size 2×2 and for fully polarimetric data it is of size 3×3. In the multitemporal case, this is calculated independently for each acquisition and summed for each class.

4. RESULTS

Results are obtained by training each method on the training part of the image illustrated in Figure 1b. For the dictionary learning method the training amounts to building the dictionary given an atom size $\sqrt{n}$, a number of temporal acquisitions (time points) Δt to include and the polarimetric mode, implicitly defining p. Training the ML classifier for a given Δt and polarimetric mode corresponds to estimating the class covariances and means, such that Eq. (5) and (6) can be evaluated for a new observation.

The classification results are shown numerically in Table 2, where the best performing representation/parameter combination is shown in bold. The numbers are classification errors, i.e., the percentage of misclassified observations in each class. The results are divided into polarimetric modes by row and number of temporal acquisitions Δt by column. The width of the dictionary atoms and the single-pixel ML classifier are included as sub-columns.

Δt	1				2				3				4			
$\sqrt{n}$	3	5	7	ML	3	5	7	ML	3	5	7	ML	3	5	7	ML
HH	0.55	0.52	0.51	0.88	0.42	0.41	0.44	0.76	0.36	0.35	0.35	0.72	0.39	0.39	0.39	0.70
VV	0.61	0.59	0.58	0.84	0.41	0.42	0.46	0.65	0.37	0.33	0.38	0.51	0.29	0.28	0.29	0.56
HHVV	0.45	0.45	0.43	0.69	0.27	0.27	0.30	0.44	0.26	0.25	0.25	0.35	0.24	0.24	0.27	**0.30**
Full	0.42	0.43	0.41	0.71	0.27	0.29	0.33	0.50	0.26	0.27	0.27	0.37	**0.22**	0.23	0.25	0.32

Table 2: Classification errors. Polarimetric mode is by row, number of temporal acquisitions by column and dictionary size by sub-column together with the single-pixel ML classifier. The minimum error for each classifier is marked in bold.

It is seen that the classification errors for the dictionary learning approach ranges from 22% to 61%, while the ML classifier has a minimum of 30% and maximum of 88%. For both classifiers the maximum classification error is obtained using single-polarization (HH or VV) and a single acquisition. The best classification is obtained by use of dual-copolarimetry for the ML classifier and full polarimetry for the dictionary classifier. Both prefer using all four temporal acquisitions.

It should be noted that while the classification errors are much lower for the dictionary approach, the computation times are much higher. They vary from approximately 20 seconds to 7.5 minutes, while the ML classifier maximally spends 7.4 seconds. This is primarily due to the high dimensional space in which the nearest neighbor search is performed.

Figures 3a–b also show the classification error for the four polarimetric modes using dictionary learning with three different atom sizes $\sqrt{n} = \{3, 5, 7\}$ in black and the ML classifier (red dots). All error plots are classification errors as a function of Δt.

The superior method for classification of this data set is clearly the dictionary learning approach. In all cases this method has a lower classification error than the ML classifier. Collectively from the results it can be inferred that the inclusion of multiple temporal acquisitions (up to at least three) improves the crop classification, which is consistent with other studies.[5] It is also apparent that the HHVV and fully polarimetric modes bring significant information to both classification methods.

The dictionary atom's spatial extent does not seem to be a parameter for classification of this particular data set as very similar results are obtained for the three sizes tested, though it seems that the largest atom of 7×7 almost never outperforms the others. Based on this, the atom should be chosen to be 3×3 to reduce the computational load.

The best classification result for the dictionary learning method was obtained using the full polarimetric information and parameters $\Delta t = 4, \sqrt{n} = 3$. The classified image using these parameters can be seen in Figure 4a. The lowest classification error for the ML classifier was obtained using the dual-copolarization (HHVV) polarimetric mode and $\Delta t = 4$. The classified image using these parameters can be seen in Figure 4b.

Comparing the two classified images it is apparent that the spatially aware dictionary method yields a spatially more coherent classification, compared to the spatially fluctuating result of the single-pixel measurement based ML classifier.

Confusion matrices for the two classification results are shown in Tables 3a–b. Both methods show large amounts of winter barley and grass classified as rye. However, the classification of grass is superior by use of the ML classifier by approximately 14%. Interestingly, rye is misclassified as winter barley and/or winter wheat, depending on the method. Inspection of Figures 4a–b reveals that it is approximately the same region in the center rye field that is difficult for the methods to classify. More interesting is perhaps the large amount of confusion into winter barley for the ML classifier, which may be attributed to the small number of training samples for this class. The classification of rye and winter wheat suffer from the ML classifier's confusion, both being reduced approximately 20%.

	Rye	Grass	W. wheat	S. barley	Peas	W. barley
Rye	**83**		14			
Grass	24	**66**				
W. wheat	11		**63**	5	13	6
S. barley				**86**		8
Peas		7			**88**	
W. barley	65					**30**

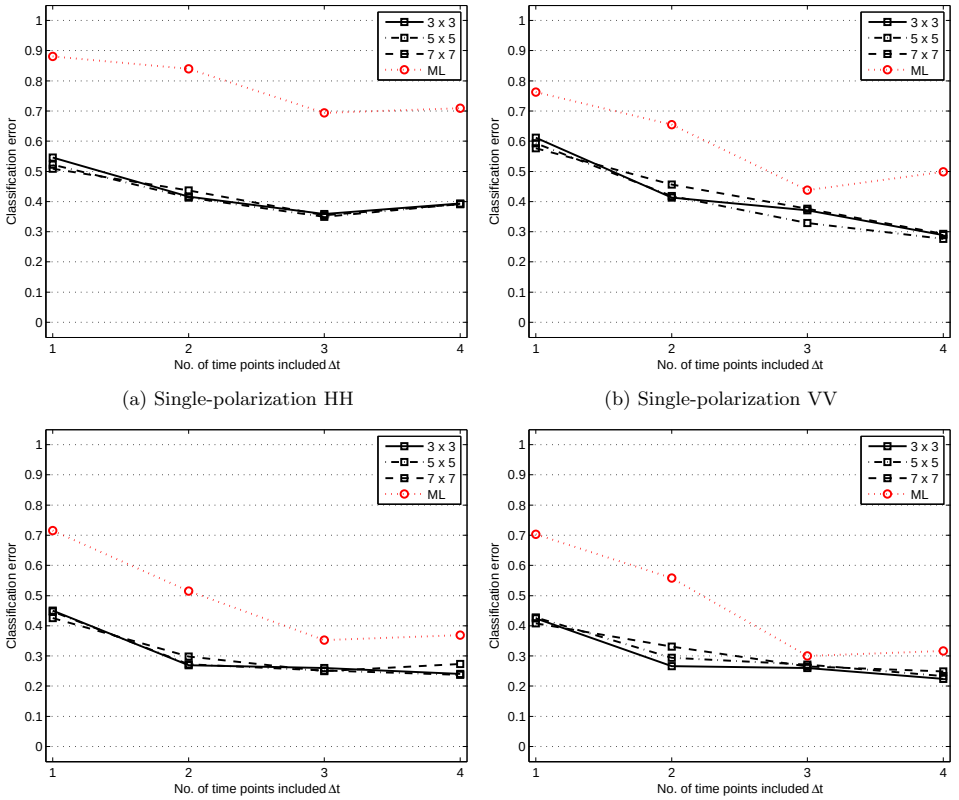
(a) Dictionary learning

	Rye	Grass	W. wheat	S. barley	Peas	W. barley
Rye	**66**		21			11
Grass	15	**80**				
W. wheat	22		**43**			30
S. barley				**88**		
Peas		9			**86**	
W. barley	50	7	12			**29**

(b) Maximum likelihood classification

Table 3: Confusion matrices belonging to the classification shown in Figures 4a–b for classification of the six classes. Numbers are percent of the row-class classified as the column-class. Errors below 5% are omitted.

A few points are worth mentioning: The large overall difference between the two methods' performances might be due to the significant difference between (1) not assuming any distribution and doing a nearest neighbor look-up, and (2) assuming a distribution and modelling each class separately. This is, however, the original forms of the two classification methods. It should be noted that the dictionary method excels in modelling the transition



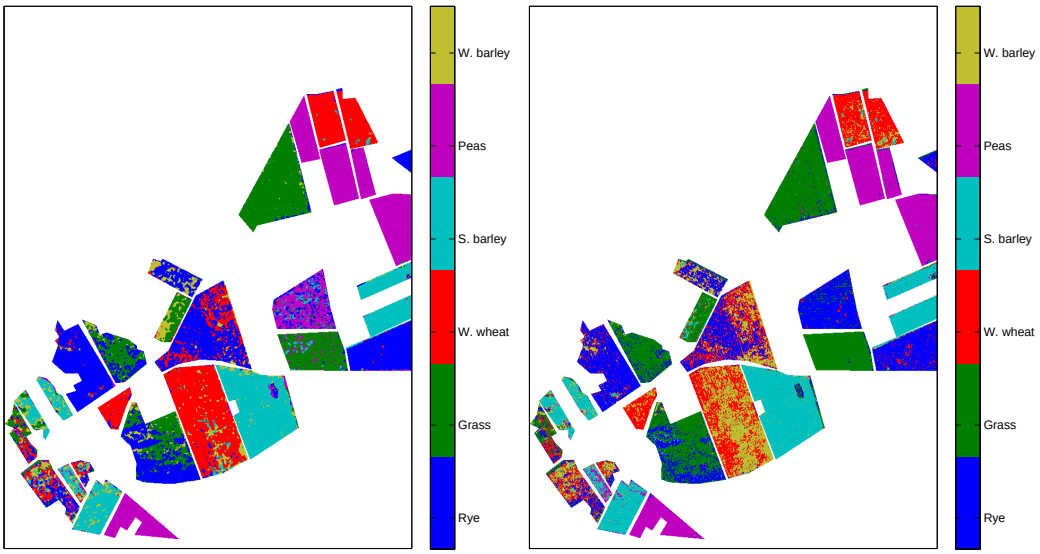
(a) Single-polarization HH

(b) Single-polarization VV

(c) Dual-copolarization HHVV

(d) Full polarimetry

Figure 3: Classification errors as a function of the number of time points included. Dictionary learning classification errors are shown in black with square markers and separate dash styles for each atom size. Maximum likelihood classification errors are shown in red with circular markers. Each plot represents one polarimetric acquisition mode.



(a) Dictionary learning (b) Maximum likelihood

Figure 4: a) Dictionary learning for minimal classification error of 26% with polarimetric mode = Full, $\Delta t = 4, \sqrt{n} = 3$. b) Maximum likelihood classification for 30% error with polarimetric mode = HHVV, $\Delta t = 4$.

between classes, which is not exploited fully here, due to the nature of the data set (all fields are separated by at least one pixel). Furthermore, it could be argued that the deterministic partition of the image data into training and test set could have an influence on the performance. While the performance in general is significantly different, the best classification for both methods is only 8% apart. Whether this is due to this particular data set, or similar performances can in fact be achieved in general for crop classification by careful choice of polarimetric mode and parameters, must be investigated in a larger scale study.

5. CONCLUSIONS

A discriminative dictionary of SAR image patches has been trained and used for classification of crops at the Foulum test site. It was found that full polarimetry SAR and inclusion of multitemporal acquisitions gave the best classification results for this method, namely a classification error of 22%. The standard maximum likelihood classifier, assuming a complex Wishart distribution for the dual-copolarimetric SAR data, achieved a classification error of 30%.

We have shown that this general approach to classification, considering contextual information and making no assumptions on distribution of the data, has a potential for crop classification in polarimetric SAR data. We have verified that inclusion of multitemporal acquisitions reduces the classification error for the classification methods evaluated here.

6. ACKNOWLEDGEMENTS

The authors would like to thank Henning Skriver, DTU Space, for making this subset of the Foulum data set available and valuable guidance.

REFERENCES

[1] Lee, J., Grunes, M., and Ainsworth, T., "Unsupervised classification using polarimetric decomposition and the complex Wishart classifier," *IEEE Transactions on Geoscience and Remote Sensing* **37**(5), 2249–2258 (1999).

[2] Liu, G., Huang, S., and Torre, A., "Bayesian classification of multi-look polarimetric SAR images with a generalized multiplicative speckle model and adaptive a priori probabilities," *International Journal of Remote Sensing* (August 2012), 37–41 (1998).

[3] Hoekman, D. and Vissers, M., "A new polarimetric classification approach evaluated for agricultural crops," *IEEE Transactions on Geoscience and Remote Sensing* **41**, 2881–2889 (Dec. 2003).

[4] Cloude, S. and Pottier, E., "An entropy based classification scheme for land applications of polarimetric SAR," *IEEE Transactions on Geoscience and Remote Sensing* **35**(1), 68–78 (1997).

[5] Skriver, H., "Crop Classification by Multitemporal C- and L-Band Single- and Dual-Polarization and Fully Polarimetric SAR," *IEEE Transactions on Geoscience and Remote Sensing* **50**(6), 2138–2149 (2012).

[6] Dahl, A. and Larsen, R., "Learning dictionaries of discriminative image patches," in [*Proceedings of the British Machine Vision Conference*], 77.1–77.11, BMVA Press (2011).

[7] Vestergaard, J. S., Dahl, A. L., Holm, P., and Larsen, R., "Pipeline for tracking neural progenitor cells," in [*Workshop on Medical Computer Vision - MICCAI 2012*], (2012).

[8] Vestergaard, J. S., *Improved nowcasting of heavy precipitation using satellite and weather radar data*, master's thesis, Technical University of Denmark (2011).

[9] Quegan, S., Le Toan, T., Skriver, H., Gomez-Dans, J., Gonzalez-Sampedro, M. C., and Hoekman, D. H., "Crop Classification with Multitemporal Polarimetric SAR Data," in [*Proceedings of the Workshop on POLinSAR - Applications of SAR Polarimetry and Polarimetric Interferometry (ESA SP-529)*], (2003).

[10] Lee, J., Grunes, M., and Pottier, E., "Quantitative comparison of classification capability: Fully polarimetric versus dual and single-polarization SAR," *IEEE Transactions on Geoscience and Remote Sensing* **39**(11), 2343–2351 (2001).

Appendices

Appendix H

B-spline registration of points to image

The reference image is denoted $\mathcal{R}$ and image values at coordinate $\mathbf{x}_i = [x_i, y_i]$ as $\mathcal{R}(\mathbf{x}_i)$. The set of N points $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{N}$ represents the grid node locations.

We model the deformation using a set of tensor B-splines. The tensor B-splines are created from separable one-dimensional splines. The set of knots $\boldsymbol{\xi}_x$ is the basis for the splines $\{B_j^x\}_{j=1}^{m_x}$ along the x-axis and similarly the set of knots $\boldsymbol{\xi}_y$ for the splines $\{B_j^y\}_{j=1}^{m_y}$ along the y-axis.

H.1 The Q matrix

First, we define the matrices $\mathbf{Q}_x$ and $\mathbf{Q}_y$ as

$$\mathbf{Q}_x = \begin{bmatrix} B_1(x_1;\boldsymbol{\xi}_x) & B_1(x_2;\boldsymbol{\xi}_x) & \dots & B_1(x_n;\boldsymbol{\xi}_x) \\ \vdots & & \ddots & \vdots \\ B_{m_x}(x_1;\boldsymbol{\xi}_x) & B_{m_x}(x_2;\boldsymbol{\xi}_x) & \dots & B_{m_x}(x_n;\boldsymbol{\xi}_x) \end{bmatrix} \tag{H.1}$$

$$\mathbf{Q}_y = \begin{bmatrix} B_1(y_1;\boldsymbol{\xi}_y) & B_1(y_2;\boldsymbol{\xi}_y) & \dots & B_1(y_n;\boldsymbol{\xi}_y) \\ \vdots & & \ddots & \vdots \\ B_{m_y}(y_1;\boldsymbol{\xi}_y) & B_{m_y}(y_2;\boldsymbol{\xi}_y) & \dots & B_{m_y}(y_n;\boldsymbol{\xi}_y) \end{bmatrix} \tag{H.2}$$

such that the (i,j)'th element of $\mathbf{Q}_x$ is the value in the j'th point of the i'th spline along the x-axis $B_i^x(x_j;\boldsymbol{\xi}_x)$.

The $N \times m_x m_y$ matrix $\mathbf{Q}'$ holds the values for each N points, of each tensor spline. Since the B-splines are separable, this matrix can be written in terms of the Khatri-Rao product $\odot$ of $\mathbf{Q}_x$ and $\mathbf{Q}_y$:

$$\mathbf{Q}' = (\mathbf{Q}_x \odot \mathbf{Q}_y)^T \,. \tag{H.3}$$

This Khatri-Rao product effectively multiplies all elements in i'th column of $\mathbf{Q}_x$ with all elements in the same column of $\mathbf{Q}_y$. Thus the matrix $\mathbf{Q}'$ has values for a single tensor spline per column, while observations are per row. The matrix referred to as $\mathbf{Q}$ in the following is defined as $\mathbf{Q} = \mathbf{I}_2 \oplus \mathbf{Q}'$.

A function for setting up $\mathbf{Q}'$ is outlined in pseudo code in Algorithm 1.

Algorithm 1 Set up matrix $\mathbf{Q}'$ of B-spline values for a set of points

Require: $\mathbf{x}$ is set of points, $\boldsymbol{\xi}_x$ is knot sequence along x-axis, $\boldsymbol{\xi}_y$ is knot sequence along y-axis.

1: **function** GETQ($\mathbf{x}, \boldsymbol{\xi}_x, \boldsymbol{\xi}_y$)
2: $Q_{ij}^x \leftarrow B_i^x(x_j;\boldsymbol{\xi}_x) \quad \forall i \in [1,N], j \in [1,m_x]$ ▷ Eq. (H.1)
3: $Q_{ij}^y \leftarrow B_i^y(y_j;\boldsymbol{\xi}_y) \quad \forall i \in [1,N], j \in [1,m_y]$ ▷ Eq. (H.2)
4: $\mathbf{Q}' \leftarrow (\mathbf{Q}_x \otimes \mathbf{Q}_y)^T$ ▷ Eq. (H.3)
5: **return** $\mathbf{Q}'$

H.2 Objective function

To optimize the grid points' correspondence with the reference image, an objective function is formulated. The objective function $\mathcal{J}(\mathbf{w})$ to be minimized

under the two-norm is

$$\mathcal{J}(\mathbf{w}) = \|\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w})\|^2 \,. \tag{H.4}$$

Here $\mathbf{w}$ is the weight vector to be determined. Consequently the i'th element w_i holds the weight of the i'th tensor spline.

Linearization of Eq. (H.4) under the norm yields:

$$\begin{aligned}
\mathcal{J}(\mathbf{w} + \Delta\mathbf{w}) &= \|\mathcal{R}(\mathbf{x} + \mathbf{Q}(\mathbf{w} + \Delta\mathbf{w}))\|^2 \\
&= \|\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w} + \mathbf{Q}\Delta\mathbf{w}))\|^2 \\
&\approx \|\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w}) + \nabla\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w})\mathbf{Q}\Delta\mathbf{w}))\|^2 \\
&= \|\mathcal{R}(\mathbf{y}) + \nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\Delta\mathbf{w}))\|^2 \quad \text{where } \mathbf{y} = \mathbf{x} + \mathbf{Q}\mathbf{w} \\
&= \mathcal{R}(\mathbf{y})^T\mathcal{R}(\mathbf{y}) + \mathcal{R}(\mathbf{y})^T(\nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\Delta\mathbf{w}) + \\
&\quad (\nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\Delta\mathbf{w})^T\left(\mathcal{R}(\mathbf{y}) + \nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\Delta\mathbf{w}\right)
\end{aligned} \tag{H.5}$$

where $\nabla\mathcal{R}(\mathbf{y})$ is the spatial gradient arranged as

$$\nabla\mathcal{R}(\mathbf{y}) = \begin{pmatrix}\mathbf{G}_x & \mathbf{G}_y\end{pmatrix} \,, \quad \mathbf{G}_x = \operatorname{diag}\left(\frac{\partial\mathcal{R}(\mathbf{y}_1)}{\partial x}, \ldots, \frac{\partial\mathcal{R}(\mathbf{y}_N)}{\partial x}\right) \,. \tag{H.6}$$

Differentiating with respect to the change in weights $\Delta\mathbf{w}$:

$$\begin{aligned}
\frac{\partial\mathcal{J}}{\partial\Delta\mathbf{w}} &= (\nabla\mathcal{R}(\mathbf{y})\mathbf{Q})^T\left[\mathcal{R}(\mathbf{y}) + \mathcal{R}(\mathbf{y}) + 2\nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\Delta\mathbf{w}\right] \\
&= 2\left(\nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\right)^T\left[\mathcal{R}(\mathbf{y}) + \nabla\mathcal{R}(\mathbf{y})\mathbf{Q}\Delta\mathbf{w}\right] \,.
\end{aligned} \tag{H.7}$$

Defining the Jacobian $\mathbf{A} = \nabla\mathcal{R}(\mathbf{y})\mathbf{Q}$ and setting the derivative equal to zero yields:

$$\begin{aligned}
\frac{\partial\mathcal{J}}{\partial\Delta\mathbf{w}} = 0 = 2\mathbf{A}^T\left(\mathcal{R}(\mathbf{y}) + \mathbf{A}\Delta\mathbf{w}\right) &\Rightarrow \\
0 = \mathbf{A}^T\mathcal{R}(\mathbf{y}) + \mathbf{A}^T\mathbf{A}\Delta\mathbf{w} &\Leftrightarrow \\
\mathbf{A}^T\mathbf{A}\Delta\mathbf{w} = -\mathbf{A}^T\mathcal{R}(\mathbf{y}) \,.
\end{aligned} \tag{H.8}$$

Also useful is the gradient in the current point $\mathbf{w}$ which is obtained by differentiation of Eq. (H.4):

$$\frac{\partial\mathcal{J}}{\partial\mathbf{w}} = 2(\nabla\mathcal{R}(\mathbf{y})\mathbf{Q})^T\mathcal{R}(\mathbf{y}) = 2A^T\mathcal{R}(\mathbf{y}) \,. \tag{H.9}$$

Adding regularization Redefining the objective function to include regularization of the weights with parameter α, such that it reads

$$\mathcal{J}(\mathbf{w}) = \|\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w})\|^2 + \alpha \mathbf{w}^T \mathbf{I} \mathbf{w} \tag{H.10}$$

with gradient

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = 2(\nabla \mathcal{R}(\mathbf{y})\mathbf{Q})^T \mathcal{R}(\mathbf{y}) + 2\alpha \mathbf{w} = 2(A^T \mathcal{R}(\mathbf{y}) + \alpha \mathbf{w}) \; . \tag{H.11}$$

Linearization under the norm and differentiation of this objective function yields:

$$\frac{\partial \mathcal{J}}{\partial \Delta \mathbf{w}} = 2\left(\mathbf{A}^T \left[\mathcal{R}(\mathbf{y}) + \mathbf{A}\mathbf{w}\right] + \alpha(\mathbf{w} + \Delta \mathbf{w})\right) \tag{H.12}$$

Setting this derivative equal to zero yields the linear system of equations:

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \Delta \mathbf{w}} = 0 &= 2\left(\mathbf{A}^T \left[\mathcal{R}(\mathbf{y}) + \mathbf{A}\Delta \mathbf{w}\right] + \alpha(\mathbf{w} + \Delta \mathbf{w})\right) \Rightarrow \\ 0 &= \mathbf{A}^T \mathcal{R}(\mathbf{y}) + \mathbf{A}^T \mathbf{A} \Delta \mathbf{w} + \alpha \mathbf{w} + \alpha \Delta \mathbf{w} \Leftrightarrow \\ \left(\mathbf{A}^T \mathbf{A} + \alpha \mathbf{I}\right) \Delta \mathbf{w} &= -\mathbf{A}\mathcal{R}(\mathbf{y}) - \alpha \mathbf{w} \; . \end{aligned} \tag{H.13}$$

Letting observations contribute individually We now allow each observation to contribute with an individual weight to the objective function. The weights are collected in a diagonal matrix $\mathbf{C}$ with elements $C_{ii} = c_i, i \in [1, N]$ where $c_i \in [0, 1]$. The objective function now reads

$$\mathcal{J}(\mathbf{w}) = \|\mathbf{C}\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w})\|^2 + \alpha \mathbf{w}^T \mathbf{I} \mathbf{w} \tag{H.14}$$

with gradient

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = 2(\mathbf{A}^T \mathbf{C}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \alpha \mathbf{w}) \; . \tag{H.15}$$

Linearization under the norm and differentiation yields:

$$\frac{\partial \mathcal{J}}{\partial \Delta \mathbf{w}} = 2\left[\mathbf{A}^T \mathbf{C}^T \mathbf{C} \left(\mathcal{R}(\mathbf{y}) + \mathbf{A}\Delta \mathbf{w}\right) + \alpha(\mathbf{w} + \Delta \mathbf{w})\right] \; . \tag{H.16}$$

Setting this derivative equal to zero results in the linear system of equations to solve for $\Delta \mathbf{w}$:

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \Delta \mathbf{w}} = 0 &= \mathbf{A}^T \mathbf{C}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \mathbf{A}^T \mathbf{C}^T \mathbf{C} \mathbf{A} \Delta \mathbf{w} + \alpha(\mathbf{w} + \Delta \mathbf{w}) \\ &= \mathbf{B}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \mathbf{B}^T \mathbf{B} \Delta \mathbf{w} + \alpha \mathbf{w} + \alpha \Delta \mathbf{w} \Leftrightarrow \\ \mathbf{B}^T \mathbf{B} \Delta \mathbf{w} + \alpha \Delta \mathbf{w} &= -\left(\mathbf{B}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \alpha \mathbf{w}\right) \Leftrightarrow \\ (\mathbf{B}^T \mathbf{B} + \alpha) \Delta \mathbf{w} &= -\left(\mathbf{B}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \alpha \mathbf{w}\right) \end{aligned} \tag{H.17}$$

where $\mathbf{B} = \mathbf{C}\mathbf{A}$.

Diffusion type regularization Defining a diffusion type regularizer $\mathcal{S}(\mathbf{u})$ as

$$\mathcal{S}(\mathbf{u}) = \sum_{\mathbf{x}_i \in \mathbf{x}} \left(\frac{\partial \mathbf{u}}{\partial x}\right)^2 + \left(\frac{\partial \mathbf{u}}{\partial y}\right)^2 \tag{H.18}$$

where $\mathbf{x}_i = [x_i, y_i]$ are the points to be registered and $\mathbf{u} = \mathbf{Q}\mathbf{w}$ is the deformation, i.e., defined for each $\mathbf{x}_i$. The deformation of a single point can be written in terms of the basis functions and the weights

$$u(\mathbf{x}_i) = \sum_{j=1}^{m_x} \sum_{k=1}^{m_y} B_j(x_i) B_k(y_i) w_{ij} \; . \tag{H.19}$$

Thus we can write the elements of the diffusion equation in terms of the B-spline derivatives

$$\begin{aligned} \frac{\partial u(\mathbf{x}_i)}{\partial x} &= \sum_{j=1}^{m_x} \sum_{k=1}^{m_y} B'_j(x_i) B_k(y_i) w_{ij} \\ \frac{\partial u(\mathbf{x}_i)}{\partial y} &= \sum_{j=1}^{m_x} \sum_{k=1}^{m_y} B_j(x_i) B'_k(y_i) w_{ij} \end{aligned} \tag{H.20}$$

where $b'_j(x_i)$ is the derivative of the j'th B-spline along the x-axis in x_i. We arrange these derivatives in a matrix of same form as in Eqs. (H.1) and (H.2)

$$\mathbf{Q}'_x = \begin{bmatrix} B'_1(x_1; \boldsymbol{\xi}_x) & B'_1(x_2; \boldsymbol{\xi}_x) & \dots & B'_1(x_n; \boldsymbol{\xi}_x) \\ \vdots & & \ddots & \vdots \\ B'_{m_x}(x_1; \boldsymbol{\xi}_x) & B'_{m_x}(x_2; \boldsymbol{\xi}_x) & \dots & B'_{m_x}(x_n; \boldsymbol{\xi}_x) \end{bmatrix} \tag{H.21}$$

$$\mathbf{Q}'_y = \begin{bmatrix} B'_1(y_1; \boldsymbol{\xi}_y) & B'_1(y_2; \boldsymbol{\xi}_y) & \dots & B'_1(y_n; \boldsymbol{\xi}_y) \\ \vdots & & \ddots & \vdots \\ B'_{m_y}(y_1; \boldsymbol{\xi}_y) & B'_{m_y}(y_2; \boldsymbol{\xi}_y) & \dots & B'_{m_y}(y_n; \boldsymbol{\xi}_y) \end{bmatrix} \; . \tag{H.22}$$

The interactions from Eqs. (H.20) can now be collected in $N \times m_x m_y$ matrices as

$$\mathbf{D}_x = (\mathbf{Q}'_x \odot \mathbf{Q}_y)^T \tag{H.23}$$

$$\mathbf{D}_y = (\mathbf{Q}_x \odot \mathbf{Q}'_y)^T \tag{H.24}$$

and the derivative of the deformation in each point is collected in two N-vectors

$$\begin{aligned} \frac{\partial u(\mathbf{x})}{\partial x} &= \mathbf{D}_x \mathbf{w} \\ \frac{\partial u(\mathbf{x})}{\partial y} &= \mathbf{D}_y \mathbf{w} \; . \end{aligned} \tag{H.25}$$

Finally the regularizer from Eq. (H.18) can be written in terms of these as

$$\begin{aligned}\mathcal{S}(\mathbf{u}) &= \sum_{\mathbf{x}_i \in \mathbf{x}} \left(\frac{\partial \mathbf{u}}{\partial x}\right)^2 + \left(\frac{\partial \mathbf{u}}{\partial y}\right)^2 \\ &= (\mathbf{D}_x \mathbf{w})^T (\mathbf{D}_x \mathbf{w}) + (\mathbf{D}_y \mathbf{w})^T (\mathbf{D}_y \mathbf{w}) \\ &= \mathbf{w}^T \mathbf{D}_x^T \mathbf{D}_x \mathbf{w} + \mathbf{w}^T \mathbf{D}_y^T \mathbf{D}_y \mathbf{w} \\ &= \mathbf{w}^T (\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) \mathbf{w} \\ &= \mathbf{w}^T \mathbf{D} \mathbf{w} \end{aligned} \tag{H.26}$$

where we have defined $\mathbf{D} = \mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y$.

Rewriting the objective function in terms of this diffusion regularizer yields

$$\mathcal{J}(\mathbf{w}) = \|\mathbf{C}\mathcal{R}(\mathbf{x} + \mathbf{Q}\mathbf{w})\|^2 + \alpha(\|\mathbf{D}_x \mathbf{w}\|^2 + \|\mathbf{D}_y \mathbf{w}\|^2) \,. \tag{H.27}$$

The gradient is

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = 2(\mathbf{A}^T \mathbf{C}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \alpha \mathbf{D} \mathbf{w}) \,. \tag{H.28}$$

The linear system of equations to be solved for the change in weights $\Delta \mathbf{w}$ is

$$(\mathbf{A}^T \mathbf{C}^T \mathbf{C} \mathbf{A} + \alpha \mathbf{D}) \Delta \mathbf{w} = -\left(\mathbf{A}^T \mathbf{C}^T \mathbf{C} \mathcal{R}(\mathbf{y}) + \alpha \mathbf{D} \mathbf{w}\right) \,. \tag{H.29}$$

Optimization of the objective functions in Eqs. (H.4), (H.10), (H.14) and (H.27) can be formulated in a Gauss-Newton type algorithm, as in Algorithm 2. A backtracking line search algorithm is implemented as described in Vester-Christensen et al. (2008).

H.3 Expanding grid algorithm

Locating all spots (hexagon centers) in an image starts from a given seed point and a bounding box, constraining the final extent of the grid nodes. From the seed point, the grid is grown outwards. Algorithm 3 summarizes the strategy, namely alternating between adding and aligning points.

Algorithm 2 Align grid of points to reference image using B-spline deformations

Require: $\mathbf{x}$ is a set of points, $\mathcal{R}$ is the reference image, α is a regularization parameter on the deformation field, $\mathbf{w}$ is initial weight vector.

function AlignGrid($\mathcal{R}, \mathbf{x}, \mathbf{Q}', \mathbf{w}, \alpha$)
  $\mathbf{Q} \leftarrow \mathbf{I}_2 \otimes \mathbf{Q}'$
  while !stop **do**
    $\mathbf{y} \leftarrow \mathbf{x} + \mathbf{Q}\mathbf{w}$
    $\mathbf{A} \leftarrow \nabla\mathcal{R}(\mathbf{y})\mathbf{Q}$
    $\Delta\mathbf{w} \leftarrow \left(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{I}\right)^{-1}\left(-\mathbf{A}^T\mathcal{R}(\mathbf{y}) - \alpha\mathbf{w}\right)$
    $\beta \leftarrow \texttt{linesearch}(\Delta\mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \beta\Delta\mathbf{w}$
  return $\mathbf{w}, \mathbf{y}$

Algorithm 3 Find hexagon centers by alternating between expanding and aligning grid. The function ExpandGrid simply adds a line of hexagon centers to the left, right, top or bottom of the existing grid. The direction of expansion depends on `k`.

Require: $\mathbf{x}$ is set of initial points, $\mathcal{R}$ is reference image, α is regularization parameter, `bb` specifies a bounding box for the grid.

function FindSpots($\mathcal{R}, \mathbf{x}$, `bb`)
  $\mathbf{w} \leftarrow \mathbf{0}, \texttt{k} \leftarrow 0$
  while #added ¿ 0 **do**
    `k++`
    $\mathbf{x}_{\text{new}} \leftarrow$ ExpandGrid(`k`,`bb`) ▷ $\mathbf{x}_{\text{new}}$: Points added.
    $\mathbf{Q}'_{\text{new}} \leftarrow$ getQ($\mathbf{x}_{\text{new}}, \boldsymbol{\xi}_x, \boldsymbol{\xi}_y$)
    $\mathbf{Q}' \leftarrow \begin{bmatrix} \mathbf{Q}' \\ \mathbf{Q}'_{\text{new}} \end{bmatrix}, \quad \mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_{\text{new}} \end{bmatrix}$
    $\mathbf{w}, \mathbf{y} \leftarrow$ AlignGrid($\mathcal{R}, \mathbf{x}, \mathbf{Q}', \mathbf{w}, \alpha$)
    #added $\leftarrow |\mathbf{x}_{\text{new}}|$ ▷ $|\cdot|$: Cardinality of set
  return $\mathbf{y}$

Appendix I

Quadratic surface fitting

A quadratic surface can be parameterized as

$$z(x, y) = ax^2 + by^2 + cxy + dx + ey + f \ . \tag{I.1}$$

Fitting a quadratic surface in an $n \times n$ window amounts to estimating the parameters $a, b, \ldots, f$ from the n^2 values. Centering the patch on (x_0, y_0) the x and y-coordinates are in relation to this center.

Observing a patch Z of size 5×5 with values

$$Z = \begin{bmatrix} z_{1,1} & z_{1,2} & z_{1,3} & z_{1,4} & z_{1,5} \\ z_{2,1} & z_{2,2} & z_{2,3} & z_{2,4} & z_{2,5} \\ z_{3,1} & z_{3,2} & z_{3,3} & z_{3,4} & z_{3,5} \\ z_{4,1} & z_{4,2} & z_{4,3} & z_{4,4} & z_{4,5} \\ z_{5,1} & z_{5,2} & z_{5,3} & z_{5,4} & z_{5,5} \end{bmatrix} = \begin{bmatrix} z_1 & z_6 & z_{11} & z_{16} & z_{21} \\ z_2 & z_7 & z_{12} & z_{17} & z_{22} \\ z_3 & z_8 & z_{13} & z_{18} & z_{23} \\ z_4 & z_9 & z_{14} & z_{19} & z_{24} \\ z_5 & z_{10} & z_{15} & z_{20} & z_{25} \end{bmatrix}$$

where two-element subscripts denote row/column indices and one-element subscripts denote linear indices.

For such a patch, the x and y-coordinates are:

$$\bar{x} = (x - x_0) = \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \quad \bar{y} = (y - y_0) = \begin{bmatrix} -2 & -2 & -2 & -2 & -2 \\ -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix} .$$

The parameters are estimated by, e.g., the Least Squares solution to the linear system of equations

$$\begin{bmatrix} z_1 \\ z_1 \\ \vdots \\ z_{n^2} \end{bmatrix} = \begin{bmatrix} \bar{x}_1^2 & \bar{y}_1^2 & \bar{x}_1\bar{y}_1 & \bar{x}_1 & \bar{y}_1 & 1 \\ \bar{x}_2^2 & \bar{y}_2^2 & \bar{x}_2\bar{y}_2 & \bar{x}_2 & \bar{y}_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{n^2}^2 & \bar{y}_{n^2}^2 & \bar{x}_{n^2}\bar{y}_{n^2} & \bar{x}_{n^2} & \bar{y}_{n^2} & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} + \epsilon \, . \tag{I.2}$$

Having obtained the parameters $a, b, \ldots, f$, its extremum $(x^\star, y^\star)$ can be found by differentiating (I.1) and setting equal to zero:

$$\frac{\partial z}{\partial x} = 0 \wedge \frac{\partial z}{\partial y} = 0 \Leftrightarrow$$
$$2ax^\star + cy^\star + d = 0 \wedge 2by^\star + cx^\star + e = 0$$

This can be formulated as a linear system of equations, which can be solved for $(x^\star, y^\star)$:

$$\begin{bmatrix} 2a & c \\ c & 2b \end{bmatrix} \begin{bmatrix} x^\star \\ y^\star \end{bmatrix} = \begin{bmatrix} -d \\ -e \end{bmatrix} \tag{I.3}$$

Appendix J

Microsatelite analysis in kernel space

Microsatellite data for a single individual consist of the number of repeats of a given microsatellite at each chromosome. As such, microsatellite data are multi-allelic. The microsatellite occurs as a molecular marker at – ideally – the same locus across individuals. It is assumed that a given repeat number occurs as a new mutation in the genome only once. Thus, if two individuals share the same repeat number they are assumed to have inherited it from a common ancestor. If the number of repeats are not the same, the allele is not shared.

Microsatellites differ from single-nucleotide polymorphisms (SNPs) in that three or more repeat number alleles can occur at each locus, where SNPs are binary markers.

The multi-allelic nature of the microsat data complicates the use of standard data analysis tools (Hansen et al., 2001). Let A be a given number of repeats and $B \neq A$ another number of repeats of a given microsat. At a bi-allelic locus, an individual can have one of the following configurations $\{A, A\}, \{A, B\}, \{B, B\}$. $\{B, A\}$ is not included, as the configuration needs to be seen as an unordered set and is thus the same as $\{A, B\}$.

J.1 A Mercer kernel for microsatellite data

Here a kernel function, valid according to Mercer (1909), is derived. Martin (2011) made a similar derivation, where the distance measure used here is referred to as the Nei-Li distance.

A simple, yet accepted (Murray, 1996), distancemeasure between two individuals' microsatellite responses is the average number of differing alleles. The set of alleles for the i'th individual $\mathcal{S}_i$ is an unordered set. Thus a listing of possible distances $d(\mathcal{S}_1, \mathcal{S}_2)$ are:

$$\begin{aligned} d(\{A,A\},\{A,A\}) &= 0 \\ d(\{A,B\},\{A,B\}) &= 0 \\ d(\{A,A\},\{A,B\}) &= 0.5 \\ d(\{A,A\},\{B,A\}) &= 0.5 \\ d(\{A,A\},\{B,B\}) &= 1\ . \end{aligned}$$

A similarity measure exactly fulfilling this is the cardinality of the intersection of the two sets divided by the average cardinality of each set. Thus a distance between the two sets can be written as

$$d(\mathcal{S}_i, \mathcal{S}_j) = 1 - \frac{2|\mathcal{S}_i \cap \mathcal{S}_j|}{|\mathcal{S}_i| + |\mathcal{S}_j|}\ . \tag{J.1}$$

The similarity measure incorporated here is also known as Sørensen's coefficient or Dice's coefficient (Dice, 1945). It is widely used in, e.g., image analysis to quantify amount of overlap between two segmentation results.

The distance measure immediately extends to multiple microsatellites per individual as an average over all microsats.

Claim 1. *The function*

$$d(\mathcal{S}_i, \mathcal{S}_j) = 1 - \frac{2|\mathcal{S}_i \cap \mathcal{S}_j|}{|\mathcal{S}_i| + |\mathcal{S}_j|}$$

of two sets $\mathcal{S}_i$ and $\mathcal{S}_j$ with $|\bar{\mathcal{S}}| \equiv |\mathcal{S}_i| = |\mathcal{S}_j|$ is a valid distance metric and the symmetric matrix $\mathbf{K} \in \mathbb{R}^{m\times m}$ with elements $K_{ij} = d(\mathcal{S}_i, \mathcal{S}_j) \quad i,j \in [1,m]$ is thus a positive semi-definite matrix.

Proof. The function d need to fulfil the four conditions

1. Non-negativity: $d(\mathcal{S}_i, \mathcal{S}_j) \geq 0$.

- Since $|\mathcal{S}_i \cap \mathcal{S}_j| \leq \min(|\mathcal{S}_i|, |\mathcal{S}_j|)$ then $\frac{2|\mathcal{S}_i \cap \mathcal{S}_j|}{|\mathcal{S}_i|+|\mathcal{S}_j|} \leq 1$ and the condition is thus fulfilled.

2. Coincidence: $d(\mathcal{S}_i, \mathcal{S}_j) = 0$ iff $S_i = S_j$

   - Since $|\mathcal{S}_i \cap \mathcal{S}_j| = \min(|\mathcal{S}_i|, |\mathcal{S}_j|) = |\bar{\mathcal{S}}|$ if and only if $\mathcal{S}_i = \mathcal{S}_j$ the condition is fulfilled. This holds due to the equal cardinality of the two sets.

3. Symmetry: $d(\mathcal{S}_i, \mathcal{S}_j) = d(\mathcal{S}_j, \mathcal{S}_i)$.

   - Trivially fulfilled from the definition.

4. Triangle inequality: $d(\mathcal{S}_i, \mathcal{S}_j) \leq d(\mathcal{S}_i, \mathcal{S}_k) + d(\mathcal{S}_j, \mathcal{S}_k)$.

   - The following must hold:

$$2 - \frac{2|\mathcal{S}_i \cap \mathcal{S}_k|}{|\mathcal{S}_i| + |\mathcal{S}_k|} - \frac{2|\mathcal{S}_j \cap \mathcal{S}_k|}{|\mathcal{S}_j| + |\mathcal{S}_k|} \geq 1 - \frac{2|\mathcal{S}_i \cap \mathcal{S}_j|}{|\mathcal{S}_i| + |\mathcal{S}_j|} \Leftrightarrow$$
$$|\mathcal{S}_i \cap \mathcal{S}_k| + |\mathcal{S}_j \cap \mathcal{S}_k| \leq |\bar{\mathcal{S}}| + |\mathcal{S}_i \cap \mathcal{S}_j| \qquad \text{(J.2)}$$

   . Decomposing into disjoint sets

$$|\mathcal{S}_i \cap \mathcal{S}_k| + |\mathcal{S}_j \cap \mathcal{S}_k| = |\mathcal{S}_i \cap \mathcal{S}_j \cap \mathcal{S}_k| + |\mathcal{S}_j \cap \mathcal{S}_k| + |(S_i \setminus (\mathcal{S}_i \cap \mathcal{S}_j)) \cap \mathcal{S}_k|$$

   and since $|\mathcal{S}_i \cap \mathcal{S}_j \cap \mathcal{S}_k| \leq |\mathcal{S}_i \cap \mathcal{S}_j|$ and $|\mathcal{S}_j \cap \mathcal{S}_k| + |(S_i \setminus (\mathcal{S}_i \cap \mathcal{S}_j)) \cap \mathcal{S}_k| \leq |\bar{\mathcal{S}}|$ this implies that (J.2) is fulfilled.

The semi-positive definiteness of $\mathbf{K}$ follows directly from the fact that it is symmetric and non-negative.

□

A set of conditions formulated by Mercer (1909) ensure that the kernel space V is a reproducing kernel Hilbert space (RKHS). The validity of Claim 1 also implies that a kernel with entries according to the distance measure in Eq. (J.1) is a valid Mercer-kernel (Mercer, 1909).

Appendix K

Reaction-diffusion mechanisms

The reaction-diffusion (R-D) equation for a one chemical system is

$$\frac{\partial \mathbf{a}}{\partial t} = f(\mathbf{a}) + D\nabla^2 \mathbf{a}\,, \tag{K.1}$$

where $\mathbf{a} = [a_1, \ldots, a_N]^T$ is the vector of concentrations of the chemical (or morphogen) in each N cells at some time t, $f(\cdot)$ describes the element wise reaction, D is the diffusion coefficient and $\nabla^2\mathbf{a}$ the Laplacian incorporating the spatial characteristics of the diffusivity, i.e., accounting for nearby contributions to the concentration of a_i due to diffusion. Two chemical systems can be modelled as

$$\frac{\partial \mathbf{a}}{\partial t} = f(\mathbf{a}, \mathbf{b}) + D_a\nabla^2 \mathbf{a} \tag{K.2}$$

$$\frac{\partial \mathbf{b}}{\partial t} = g(\mathbf{a}, \mathbf{b}) + D_b\nabla^2 \mathbf{b} \tag{K.3}$$

where b is concentration for the second chemical (Murray, 2002). Turing (1952) argued that this set of equations can be used to drive pattern formation: in the abscence of diffusion ($D_a = D_b = 0$) the concentrations will stabilize, but when $D_a \neq D_b$ and other certain conditions are fulfilled spatially inhomogeneous patterns can emerge due to diffusion driven instability (Murray, 2002). Gierer and Meinhardt (1972) has shown that local self-enhancement and long-range

inhibition in the R-D system is the driving force of pattern formation and models have been derived on that basis for a wide range of biological systems, (see e.g., Bard, 1981, Meinhardt, 1993, Koch and Meinhardt, 1994, Meinhardt, 1999, Shoji et al., 2003, Kondo and Miura, 2010, Allen et al., 2013). Turk (1991) coins the requirements for pattern formation:

> The key to pattern formation based on reaction-diffusion is that an initial small amount of variation in the chemical concentrations can cause the system to be unstable initially and to be driven to a stable state in which the concentrations of a and b vary across the surface.

Let us consider the reaction equations proposed by Turk (1991):

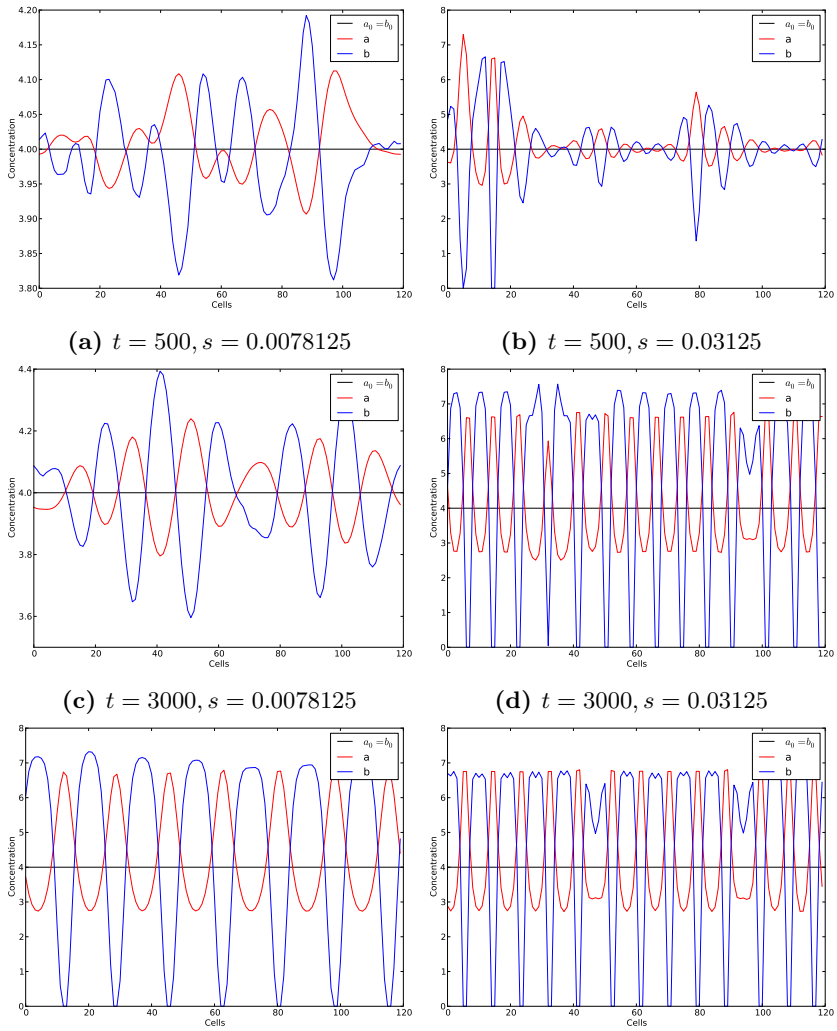
$$f(a_i, b_i) = s(16 - a_i b_i) \tag{K.4}$$
$$g(a_i, b_i) = s(a_i b_i - b_i - \beta_i)\ , \tag{K.5}$$

where $\beta_i \sim \mathcal{N}(a_i b_i - b_i, \sigma_p^2)$ is a slight random perturbation to the initial concentrations, s is a scaling factor related to the size of the domain. Turk proposes the discrete approximation of the Laplacian in 1D by $\nabla^2 x_i = x_{i+1} + x_{i-1} - 2a_i$ where $\mathbf{x} \in \mathbb{R}^N$ and on a 2D regular grid by $\nabla^2 X_{ij} = X_{i+1,j} + X_{i-1,j} + X_{i,j+1} + X_{i,j-1} - 4X_{ij}$ where $\mathbf{X} \in \mathbb{R}^{M \times N}$. This discretization is conveniently implementable as a convolution and assuming periodic boundary conditions, i.e., $x_{N+1} = x_1$, reduces the boundary artefacts.

Figure K.1 shows simulation results in a 1D domain for $N = 120$ cells at different t and different scaling factors. It is seen how the scaling factor influences the periodicity of the pattern and how a less localized mechanism (small scaling factor) converges much slower than a very localized (large scaling factor) mechanism. This is due to the diffusion "trickles off" faster in the latter case.

Similarly a 2D simulation can be seen in Figure K.2, where chemical **a** is spread over a domain of size $M \times N = 120 \times 120$. The concentration is shown on a false color scale and the binary image show areas in red where $\mathbf{a} > a_0$. Chemical **b** is not shown as the pattern is similar in nature, though complimentary like in the 1D case.

The diffusivity of each chemical can influence the final patterning in remarkable ways. In Figures K.3a–K.3d two different values of D_b have been used, creating either maze-like or dotty patterns. In Figures K.3e–K.3f the value of D_b has been mapped linearly between these two values as a function of the y-axis position. This elucidates one of the ways of making controllable patterns, namely by spatial variation of one or more of the parameters. Another way would be to introduce so-called prepatterning (see e.g., Meinhardt and Meinhardt, 1982,

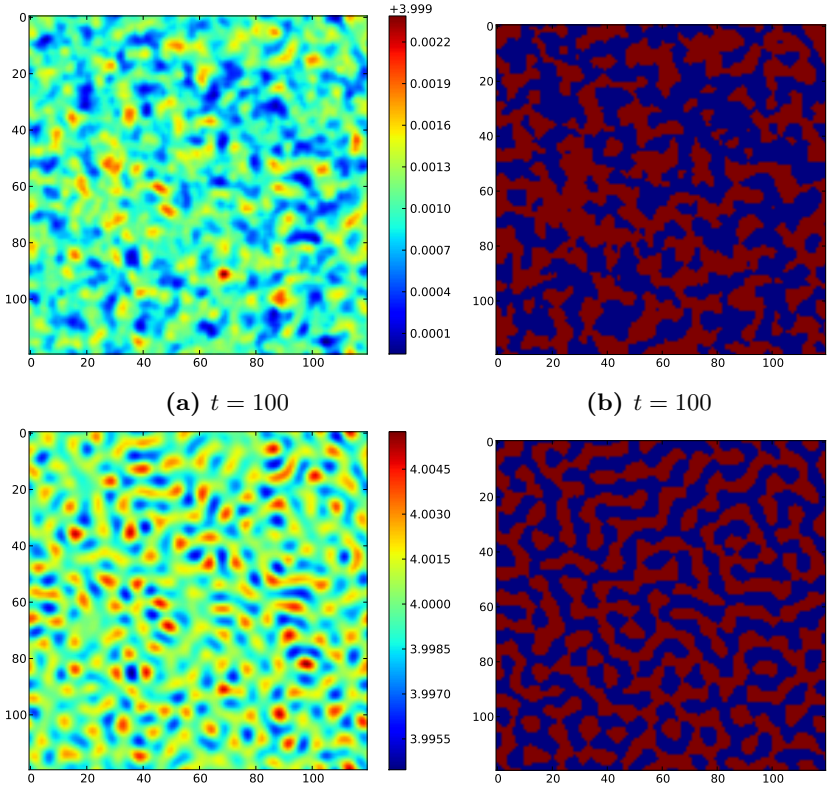


(a) $t = 500, s = 0.0078125$ **(b)** $t = 500, s = 0.03125$

(c) $t = 3000, s = 0.0078125$ **(d)** $t = 3000, s = 0.03125$

(e) $t = 10000, s = 0.0078125$ **(f)** $t = 10000, s = 0.03125$

Figure K.1: One-dimensional reaction-diffusion simulations for a two-chemical system with $N = 120$ cells, initial values $a_0 = b_0 = 4$, diffusion coefficients $D_a = D_a = 0.25, D_b = 0.0625$ and initial perturbation $\sigma_p = 0.05$. Scaling factors $s = \{0.0078125, 0.03125\}$ are used and the simulation is stopped after $t = \{500, 3000, 10000\}$ time steps.



(a) $t = 100$ **(b)** $t = 100$

(c) $t = 900$ **(d)** $t = 900$

Figure K.2: Two-dimensional reaction-diffusion simulations for a two-chemical system with 120×120 cells in a grid, initial values $a_0 = b_0 = 4$, diffusion coefficients $D_a = 0.25, D_b = 0.0625$, initial perturbation $\sigma_p = 0.001$, scaling factor $s = 0.03125$ are used and the simulation is stopped after $t = \{100, 900\}$ time steps. The left column shows concentrations of $\mathbf{a}$ and the right column shows $\mathbf{a} > 0$.

Turk, 1991, Kondo and Miura, 2010), where the pattern is laid out in a series of consecutive simulations, freezing parts of the domain along the way. Yet another way would be implementation of specific boundaries or specialized domains, or as suggested by Shoji et al. (2003) anisotropic diffusion.

Anisotropic diffusion modulates the diffusion by introducing an anisotropy magnitude $\delta_a \in]-1, 1[$ for one of the chemicals. The R-D equations in Eqs. (K.2) and (K.3) now includes an anisotropy function $\alpha(\cdot)$ such that

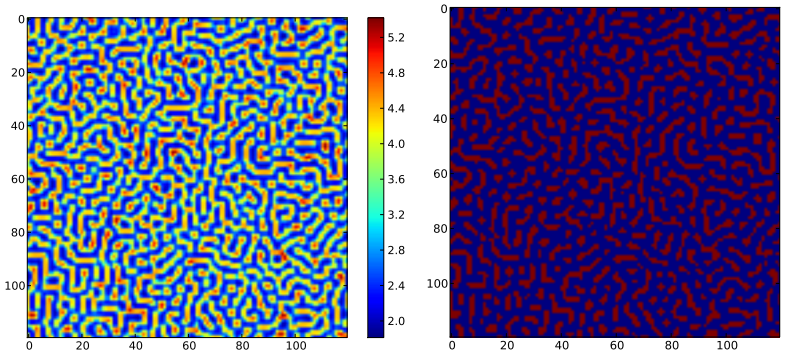
$$\frac{\partial \mathbf{a}}{\partial t} = f(\mathbf{a}, \mathbf{b}) + D_a \alpha(\theta) \nabla^2 \mathbf{a} \tag{K.6}$$

$$\frac{\partial \mathbf{b}}{\partial t} = g(\mathbf{a}, \mathbf{b}) + D_b \nabla^2 \mathbf{b} \,, \tag{K.7}$$

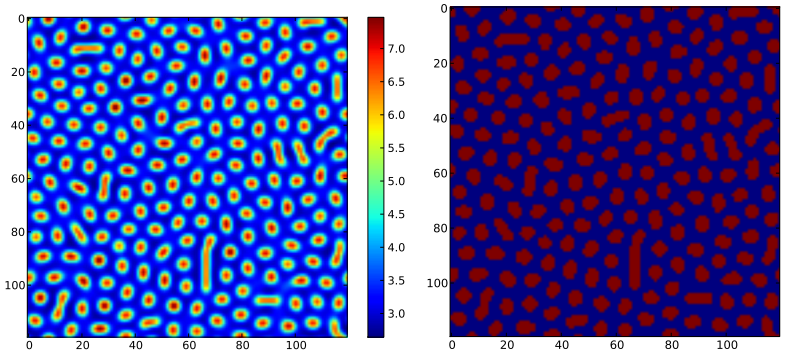
where

$$\alpha(\theta) = \frac{1}{\sqrt{1 - \delta_a \cos 2\theta}} \tag{K.8}$$

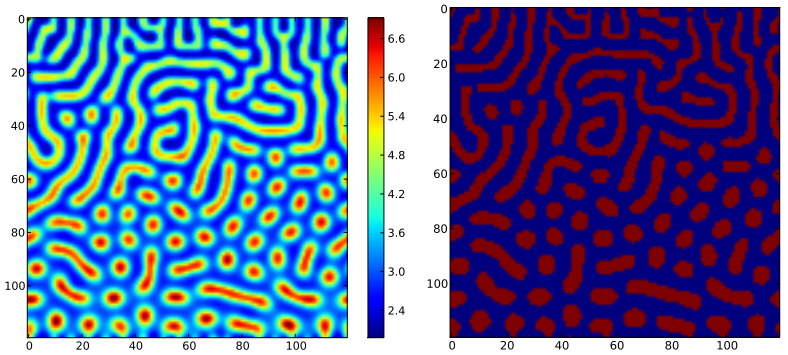
and θ is the angle to the neighboring cell currently considered. For a grid where diffusion is approximated by using only the horizontal and vertical neighbors $\theta = \{0, \frac{\pi}{2}\}$ respectively. Here we have chosen to maintain D_a as a separate term to make it easier to compare with previous experiments. It is seen how $\delta_a = 0$ reduces the model to the original in Eqs. (K.2)–(K.3) and thus corresponds to no anisotropy, $\delta_a \to -1$ results in strong diffusivity along the x-axis and $\delta_a \to 1$ results in strong diffusivity along the y-axis. This is confirmed by the simulations shown in Figure K.4.



(a) $D_b = 0.0125$ **(b)** $D_b = 0.0125$

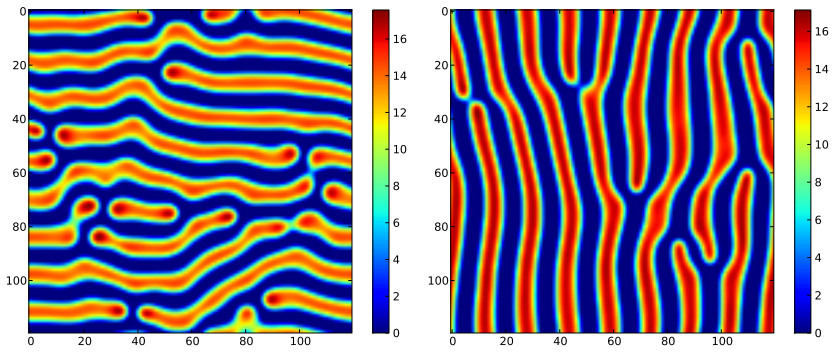


(c) $D_b = 0.05$ **(d)** $D_b = 0.05$



(e) $D_b \propto y$ **(f)** $D_b \propto y$

Figure K.3: (a)–(d) Two-dimensional reaction-diffusion simulations for a two-chemical system with same parameters as in Figure K.2, except the diffusivity D_b is different between the two simulations and 5000 time steps have been simulated. (e)–(f) Diffusivity coefficient D_b is stretched linearly between 0.0125 and 0.5 from top to bottom. Here a scaling factor of $s = 0.015625$ was used and 10000 time steps were simulated.



(a) $\delta_a = -0.95$ **(b)** $\delta_a = 0.95$

Figure K.4: Concentration of chemical b after 10000 time steps. Two different values for the anisotropic diffusion magnitude δ_a are used.

Bibliography

Abrahamsen, T. *Kernel Methods for De-noising with Neuroimaging Application.* Master's thesis, Technical University of Denmark, 2009.

Allen, W. L., Baddeley, R., Scott-Samuel, N. E., and Cuthill, I. C. The evolution and function of pattern diversity in snakes. *Behavioral Ecology*, 24(5):1237–1250, 2013. ISSN 1045-2249. doi:10.1093/beheco/art058.

Amit, Y. and Kong, a. Graphical templates for model registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):225–236, 1996. ISSN 01628828. doi:10.1109/34.485529.

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis, 2nd Edition.* Wiley-Interscience, 2 edition, 1984. ISBN 0471889873.

Astola, J. Entropy correlation coefficient, a measure of statistical dependence for categorized data. *Proc. Univ. Vaasa, Discussion Papers*, 1982.

Avidan, S. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, 2:1491–1498, 2006. doi: 10.1109/CVPR.2006.119.

Bard, J. B. A model for generating aspects of zebra and other mammalian coat patterns. *Journal of Theoretical Biology*, 93(2):363–85, 1981. ISSN 0022-5193.

Bay, H., Tuytelaars, T., and Gool, L. V. SURF: Speeded up robust features. *Computer Vision–ECCV 2006*, 2006.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373—-1396, 2003.

Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–59, 1995. ISSN 0899-7667.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 20 edition, 2007. ISBN 0387310738.

Bosch, A., Zisserman, A., and Munoz, X. Representing shape with a spatial pyramid kernel. *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*, pages 401–408, 2007. doi:10.1145/1282280.1282340.

Botev, Z. I., Grotowski, J. F., and Kroese, D. P. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010. ISSN 0090-5364. doi:10.1214/10-AOS799.

Boykov, Y. and Kolmogorov, V. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Transactions on Pattern and Machine Intelligence*, 26(9):1124–1137, 2004. ISSN 0162-8828. doi:http://dx.doi.org/10.1109/TPAMI.2004.60.

Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

Buzug, T. *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*, chapter 2.6. Springer, 2008. ISBN 9783540394082.

Cai, D., He, X., and Han, J. Efficient kernel discriminant analysis via spectral regression. In *Seventh IEEE International Conference on Data Mining. ICDM 2007.*, August. 2007a.

Cai, D., He, X., and Han, J. Spectral Regression for Efficient Regularized Subspace Learning. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007b. doi:10.1109/ICCV.2007.4408855.

Carstensen, J. *Description and simulation of visual texture*. Ph.d. thesis, Technical University of Denmark, 1992.

Carstensen, J. M. An Active Lattice Model in a Bayesian Framework. *Computer Vision and Image Understanding*, 63(2):380–387, 1996. ISSN 10773142. doi:10.1006/cviu.1996.0027.

Chan, K., Lee, T.-W., and Sejnowski, T. J. Variational Bayesian Learning of ICA with Missing Data. *Neural Computation*, 15(8):1991–2011, 2003. ISSN 0899-7667. doi:10.1162/08997660360675116.

Cherian, A., Mairal, J., Alahari, K., and Schmid, C. Mixing Body-Part Sequences for Human Pose Estimation. In *CVPR 2014-IEEE Conference on Computer Vision & Pattern Recognition*, volume 1. 2014.

Clemmensen, L. B., Hastie, T., Witten, D., and Ersbøll, B. Sparse discriminant analysis. *Technometrics*, 53(4):37–41, 2011. doi:10.1198/TECH.2011.08118.

Cloude, S. and Pottier, E. An entropy based classification scheme for land applications of polarimetric SAR. *IEEE Transactions on Geoscience and Remote Sensing*, 35(1):68–78, 1997.

Conradsen, K., a.a. Nielsen, Schou, J., and Skriver, H. A test statistic in the complex wishart distribution and its application to change detection in polarimetric SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 41(1):4–19, 2003. ISSN 0196-2892. doi:10.1109/TGRS.2002.808066.

Cootes, T., Edwards, G., and Taylor, C. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. Active shape models-their training and application. *Computer vision and image . . .*, 61(1):38—-59, 1995.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. ISBN 0471241954.

Cristinacce, D. and Cootes, T. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41:3054–3067, 2008. doi:10.1016/j.patcog.2008.01.024.

Dahl, A. and Larsen, R. Learning dictionaries of discriminative image patches, 2011. doi:10.5244/C.25.77.

Dalal, N. and Triggs, B. Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893, 2005. doi:10.1109/CVPR.2005.177.

De Bie, T. and De Moor, B. On two classes of alternatives to canonical correlation analysis, using mutual information and oblique projections. In *Proc. of the 23rd symposium on information theory in the Benelux (ITB)*. 2002.

Dice, L. R. Measures of the amount of ecologic association between species. *Ecology*, 26(3):pp. 297–302, 1945. ISSN 00129658.

Donoho, D. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. ISSN 0018-9448. doi:10.1109/TIT.2006.871582.

Efron, B. Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, 99(465):96–104, 2004. ISSN 0162-1459. doi: 10.1198/016214504000000089.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Elad, M. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010. ISBN 9781441970107.

Fisher, R. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179—-188, 1936.

Geman, S. and Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–41, 1984. ISSN 0162-8828.

Gershenfeld, N. *The Nature of Mathematical Modelling*. Cambridge University Press, 1999.

Gierer, A. and Meinhardt, H. A theory of biological pattern formation. *Kybernetik*, 12(1):30–39, 1972.

Gower, J. C. Generalized Procrustes analysis. *Psychometrika*, 40(1), 1975.

Green, P. J. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711, 1995. ISSN 00063444. doi:10.2307/2337340.

Green, A. A., Berman, M., Switzer, P., and Craig, M. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *Geoscience and Remote Sensing, IEEE Transactions on*, 26(1):65–74, 1988. ISSN 01962892. doi:10.1109/36.3001.

Gudbjartsson, H. and Patz, S. The rician distribution of noisy mri data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995. ISSN 1522-2594. doi: 10.1002/mrm.1910340618.

Güdükbay, U., Özgüç, B., and Tokad, Y. A spring force formulation for elastically deformable models. *Computers & Graphics*, 21(3):335–346, 1997.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, 2:1735–1742, 2006. doi:10.1109/CVPR.2006.100.

Hammersley, J. and Clifford, P. Markov fields on finite graphs and lattices. 1971.

Hansen, M. M., Kenchington, E., and Nielsen, E. E. Assigning individual fish to populations using microsatellite DNA markers. *Fish and Fisheries*, 2(2):93–112, 2001. ISSN 1467-2960. doi:10.1046/j.1467-2960.2001.00043.x.

Harris, C. Geometry from visual motion. In A. Blake and A. Yuille (editors), *Active Vision*, pages 263–284. MIT Press, Cambridge, MA, USA, 1993. ISBN 0-262-02351-2.

Hartelius, K. and Carstensen, J. Bayesian grid matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):162–173, 2003.

Hassner, M. and Sklansky, J. The use of Markov random fields as models of texture. *Computer Graphics and Image Processing*, 1980.

Hastie, T., Buja, A., and Tibshirani, R. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

Hastie, T. and Tibshirani, R. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological).*, 58(1):155–176, 1996.

Hastie, T., Tibshirani, R., and Friedman, J. H. *The Elements of Statistical Learning*. Springer, corrected edition, 2003. ISBN 0387952845.

He, X. and Niyogi, P. Locality preserving projections. In *NIPS*, pages 234—-241. 2003.

Hinton, G., Osindero, S., and Teh, Y. A fast learning algorithm for deep belief nets. *Neural computation*, 1554:1527–1554, 2006.

Hinton, G. E., Welling, M., Teh, Y. W., and Osindero, S. A new view of ICA. In *Int. Conf. on Independent Component Analysis and Blind Source Separation*, 5. 2001.

Hoekman, D. and Vissers, M. A new polarimetric classfication approach evaluated for agricultural crops. *IEEE Transactions on Geoscience and Remote Sensing*, 41(12):2881–2889, 2003. ISSN 0196-2892. doi:10.1109/TGRS.2003.817795.

Hotelling, H. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. doi:10.1093/biomet/28.3-4.321.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent component analysis*. Adaptive and learning systems for signal processing, communications, and control. J. Wiley, 2001. ISBN 9780471405405.

Ising, E. A contribution to the theory of ferromagnetism. *Z. Phys*, 31:253–258, 1925.

Jain, A. and Lakshmanan, S. Object matching using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):267–278, 1996. ISSN 01628828. doi:10.1109/34.485555.

Jain, A. K., Zhong, Y., and Dubuisson-Jolly, M.-P. Deformable template models: A review. *Signal Processing*, 71(2):109–129, 1998. ISSN 01651684. doi:10.1016/S0165-1684(98)00139-X.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., and LeCun, Y. What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009. doi:10.1109/ICCV.2009.5459469.

Jenssen, R. Kernel entropy component analysis. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):847–60, 2010. ISSN 1939-3539. doi:10.1109/TPAMI.2009.100.

Jolliffe, I. T. *Principal component analysis*. Springer New York, 2002.

Jones, M. and Marron, J. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical*, 91(433):401–407, 1996.

Karasuyama, M. and Sugiyama, M. Canonical dependency analysis based on squared-loss mutual information. *Neural Networks*, 34:46–55, 2012. ISSN 0893-6080. doi:10.1016/j.neunet.2012.06.009.

Kittler, J. and Föglein, J. Contextual classification of multispectral pixel data. *Image and Vision Computing*, 2(1):13–29, 1984.

Kling, J., Vestergaard, J. S., Dahl, A. L., Hansen, T. W., Larsen, R., and Wagner, J. B. Automated structure detection in hrtem images: An example with graphene. *Microscopy and Microanalysis*, 19(S2):802–803, 2013.

Koch, A. and Meinhardt, H. Biological pattern formation: from basic mechanisms to complex structures. *Reviews of Modern Physics*, 66(4):1481, 1994.

Koenderink, J. J. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984.

Koenderink, J. and Doorn, A. V. The structure of locally orderless images. *International Journal of Computer Vision*, 1999.

Koenderink, J. J. and van Doorn, A. J. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–564, 1992. ISSN 02628856. doi:10.1016/0262-8856(92)90076-F.

Kondo, S. and Miura, T. Reaction-diffusion model as a framework for understanding biological pattern formation. *Science (New York, N.Y.)*, 329(5999):1616–20, 2010. ISSN 1095-9203. doi:10.1126/science.1179047.

Kubale, M. *Graph colorings*, volume 349. American Mathematical Soc., 2004. ISBN 0821834584. doi:10.1016/j.tcs.2005.09.025.

Larsen, A. B. L. *An in-depth study of local image descriptors and their performance*. Ph.D. thesis, University of Copenhagen, 2012.

Larsen, A. B. L., Vestergaard, J. S., and Larsen, R. HEp-2 cell classification using shape index histograms with donut-shaped spatial pooling. *IEEE transactions on medical imaging*, 0062(c):1–8, 2014. ISSN 1558-254X. doi: 10.1109/TMI.2014.2318434.

Li, K. and Kanade, T. Nonnegative mixed-norm preconditioning for microscopy image segmentation. *Information Processing in Medical Imaging*, 21:362–73, 2009. ISSN 1011-2499.

Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., and America, N. E. C. L. Large-scale Image Classification : Fast Feature Extraction and SVM Training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, October, pages 1689–1696. 2011.

Lindeberg, T. *Scale-space theory in computer vision*. Springer, 1993.

Lindeberg, T. Scale-space: A framework for handling image structures at multiple scales. *CERN European Organization for Nuclear Research - Reports*, pages 27–38, 1996. ISSN 00078328.

Lowe, D. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2, 1999. doi:10.1109/ICCV.1999.790410.

Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. ISSN 0920-5691. doi:10.1023/B:VISI.0000029664.99615.94.

Mairal, J., Bach, F., and Ponce, J. Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):791–804, 2012. ISSN 1939-3539. doi:10.1109/TPAMI.2011.156.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online Learning for Matrix Factorization and Sparse Coding. *Journal ofMachine Learning Research*, 11:19–60, 2010.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 0, pages 1–8. IEEE Computer Society, Los Alamitos, CA, USA, 2008a. ISBN 978-1-4244-2242-5. doi:10.1109/CVPR.2008.4587652.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. Supervised dictionary learning. *Defense Technical Information Center*, 2008b.

Mairal, J., Sapiro, G., and Elad, M. Multiscale sparse image representationwith learned dictionaries. *Image Processing, 2007. ICIP 2007. . . .*, pages 105–108, 2007.

Martin, F. An application of kernel methods to variety identification based on SSR markers genetic fingerprinting. *BMC Bioinformatics*, 12(1):177, 2011. ISSN 1471-2105. doi:10.1186/1471-2105-12-177.

Meinhardt, H. A model for pattern formation of hypostome, tentacles, and foot in hydra: how to form structures close to each other, how to form them at a distance. *Developmental biology*, 157(2):321–333, 1993.

Meinhardt, H. Orientation of chemotactic cells and growth cones: models and mechanisms. *Journal of Cell Science*, 112(17):2867–2874, 1999.

Meinhardt, H. and Meinhardt, H. *Models of biological pattern formation*, volume 6. Academic Press London, 1982.

Mercer, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, 209(1909):415–446, 1909.

Mika, S. and Ratsch, G. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48. 1999. ISBN 078035673X. ISSN 2168-2275. doi:10.1109/TCYB.2013.2273355.

Mika, S., Rätsch, G., Weston, J., and Schölkopf, B. Invariant Feature Extraction and Classification in Kernel Spaces. In *NIPS*, volume 89, pages 526—-532. 1999.

Mikolajczyk, K. and Schmid, C. Performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–30, 2005. ISSN 0162-8828. doi:10.1109/TPAMI.2005.188.

Muller, K., Mika, S., and Ratsch, G. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–202, 2001.

Murray, B. W. The estimation of genetic distance and population substructure from microsatellite allele frequency data. `http://helix.biology.mcmaster.ca/brent/brent.html`, 1996. Accessed: 2013-11-18.

Murray, J. *Mathematical biology*. Springer, 3rd editio edition, 2002. ISBN 0387952284.

Nelder, J. A. and Mead, R. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965. doi:10.1093/comjnl/7.4.308.

Nielsen, A. A. Multiset Canonical Correlations Analysis and Multispectral, Truly Multi-temporal Remote Sensing Data. *IEEE Transactions on Image Processing*, 11(3):293–305, 2002. doi:10.1109/83.988962.

Nielsen, A. A. Kernel Maximum Autocorrelation Factor and Minimum Noise Fraction Transformations. *IEEE Transactions on Image Processing*, 20(3):612–624, 2011. ISSN 1057-7149. doi:10.1109/TIP.2010.2076296.

Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Potts, R. B. Some generalized order-disorder transformations. In *Proceedings of the Cambridge Philosophical Society*, volume 48, pages 106–109. Cambridge Univ Press, 1952.

Ripley, B. D. *Statistical inference for spatial processes*. Cambridge university press, 1991.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science (New York, N.Y.)*, 290(5500):2323–6, 2000. ISSN 0036-8075. doi:10.1126/science.290.5500.2323.

Salakhutdinov, R. and Hinton, G. A Better Way to Pretrain Deep Boltzmann Machines. In *NIPS*, 3, pages 2456—-2464. 2012.

Schölkopf, B. and Smola, A. J. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Schölkopf, B., Smola, A., and Müller, K. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299—-1319, 1998.

Scott, D. W. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979. doi:10.1093/biomet/66.3.605.

Shannon, C. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. ISBN 0521813972.

Sheather, S. and Jones, M. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690, 1991.

Shen, H. and Huang, J. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99:1015–1034, 2008. doi:10.1016/j.jmva.2007.06.007.

Shoji, H., Mochizuki, A., Iwasa, Y., Hirata, M., Watanabe, T., Hioki, S., and Kondo, S. Origin of directionality in the fish stripe pattern. *Developmental dynamics : an official publication of the American Association of Anatomists*, 226(4):627–33, 2003. ISSN 1058-8388. doi:10.1002/dvdy.10277.

Shwartz, S., Zibulevsky, M., and Schechner, Y. Fast kernel entropy estimation and optimization. *Signal Processing*, 85(5):1045–1058, 2005. ISSN 01651684. doi:10.1016/j.sigpro.2004.11.022.

Silverman, B. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

Skriver, H. Crop Classification by Multitemporal C- and L-Band Single- and Dual-Polarization and Fully Polarimetric SAR. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2138–2149, 2012.

Song, Y., Nie, F., Zhang, C., and Xiang, S. A unified framework for semi-supervised dimensionality reduction. *Pattern Recognition*, 41(9):2789–2799, 2008. ISSN 00313203. doi:10.1016/j.patcog.2008.01.001.

Switzer, P. and Green, A. A. Min/max autocorrelation factors for multivariate spatial imagery. *Computer Science and Statistics: The Interface (L. Billard, Ed.)*, page 16, 1984.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–23, 2000. ISSN 0036-8075. doi:10.1126/science.290.5500.2319.

Terrell, G. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410):470–477, 1990.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Tola, E., Lepetit, V., and Fua, P. DAISY: an efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–30, 2010. ISSN 1939-3539. doi:10.1109/TPAMI.2009.77.

Turing, A. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

Turk, G. Generating textures on arbitrary surfaces using reaction-diffusion. *ACM SIGGRAPH Computer Graphics*, 25(4):289–298, 1991. ISSN 00978930. doi:10.1145/127719.122749.

Vester-Christensen, M., Larsen, R., and Christensen, L. *Image registration and optimization in the Virtual Slaughterhouse*. Ph.D. thesis, Technical University of Denmark, 2008.

Vestergaard, J. S., Dahl, A. L., Holm, P., and Larsen, R. Dynamically constrained pipeline for tracking neural progenitor cells. In *Proceedings of SPIE Medical Imaging*, volume 8676, pages 86760B–86760B–12. International Society for Optics and Photonics, 2013a. doi:10.1117/12.2006996.

Vestergaard, J. S., Dahl, A. L., Holm, P., and Larsen, R. Pipeline for tracking neural progenitor cells. In B. Menze, G. Langs, L. Lu, A. Montillo, Z. Tu, and A. Criminisi (editors), *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, volume 7766 of *Lecture Notes in Computer Science*, pages 155–164. Springer Berlin Heidelberg, 2013b. ISBN 978-3-642-36619-2. doi:10.1007/978-3-642-36620-8_16.

Vestergaard, J. S. and Nielsen, A. A. Automated invariant alignment to improve canonical variates in image fusion of satellite and weather radar data. *Journal of Applied Meteorology and Climatology*, 2012.

Veta, M. et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *submitted to Journal of Medical Image Analysis*, 2014.

Vinh, N., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

Wackernagel, H. *Multivariate Geostatistics*. Springer, 1995.

Wang, S., Zhang, L., Liang, Y., and Pan, Q. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216–2223, 2012. doi:10.1109/CVPR.2012.6247930.

Welsh, D. and Powell, M. An upper bound for the chromatic number of a graph and its application to timetabling problems. *The Computer Journal*, 10(1):85–86, 1967.

Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Wu, F.-Y. The potts model. *Reviews of modern physics*, 54(1):235, 1982.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007. ISSN 0162-8828. doi:10.1109/TPAMI.2007.12.

Yang, J. and Wright, J. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

Yao, Y. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Springer, 2003.

Yin, X. Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91:161–176, 2004. doi:10.1016/S0047-259X(03)00129-5.

Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics*, pages 1–30, 2006.